Elements of Survey Sampling

Kluwer Texts in the Mathematical Sciences

VOLUME 15

A Graduate-Level Book Series

The titles published in this series are listed at the end of this volume.

Elements of Survey Sampling

by

Ravindra Singh

and

Naurang Singh Mangat

Punjab Agricultural University,
Ludhiana, Punjab, India

SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-90-481-4703-8 ISBN 978-94-017-1404-4 (eBook)
DOI 10.1007/978-94-017-1404-4

Printed on acid-free paper

TO THE MEMORY OF OUR PARENTS

Gayatri Devi and Narendra Singh
RS

Gurdial Kaur and Hernam Singh
NSM

Table of Contents

PREFACE xiii

1. COLLECTION OF SURVEY DATA

1.1	Need for statistical data	1
1.2	Types of data	1
1.3	Methods of collecting primary data	2
1.4	Framing of questionnaire/schedule	3
1.5	Some technical terms	4
1.6	Need for a sample	5
1.7	Sampling procedures	6
1.8	With and without replacement sampling	7
1.9	Planning and execution of sample surveys	9
	Let us do	11

2. ELEMENTARY CONCEPTS

2.1	Introduction	14
2.2	Statistical preliminaries	14
2.3	Estimator and its sampling distribution	16
2.4	Unbiased estimator	19
2.5	Measures of error	21
2.6	Confidence intervals	23
2.7	Sample size determination	26
2.8	Sampling and nonsampling errors	27
	Let us do	28

3. SIMPLE RANDOM SAMPLING

3.1	What is simple random sampling ?	30
3.2	How to draw a simple random sample ?	30
3.3	Estimation of population mean/total	33
3.4	Estimation of mean/total using distinct units	48
3.5	Determining sample size for estimating population mean/total	50
3.6	Estimation of population proportion	53
3.7	Sample size for estimation of proportion	55
3.8	Estimation of proportion using inverse sampling	56

3.9 Estimation over subpopulations 58
3.10 Some further remarks 62
Let us do 63

4. SAMPLING WITH VARYING PROBABILITIES

4.1 Introduction 67
4.2 Methods of selecting a PPS sample 67
4.3 Estimation in PPSWR sampling 70
4.4 Relative efficiency of PPSWR estimator 72
4.5 Determining sample size for estimating population mean/total 76
4.6 Sampling with PPS without replacement 77
4.7 Des Raj's ordered estimator 78
4.8 Murthy's unordered estimator 82
4.9 Horvitz-Thompson estimator 84
4.10 Sen-Midzuno method 86
4.11 Random group method 91
4.12 Relative efficiency of RHC estimator 95
4.13 Some further remarks 97
Let us do 98

5. STRATIFIED SAMPLING

5.1 Introduction 102
5.2 Notations 104
5.3 Estimation of mean and total using simple random sampling 104
5.4 Allocation of sample size 108
5.5 Relative efficiency of stratified estimator 123
5.6 Estimation of population proportion 129
5.7 Construction of strata 132
5.8 Poststratification 136
5.9 Some further remarks 138
Let us do 140

6. SYSTEMATIC SAMPLING

6.1 Linear systematic sampling 145
6.2 Circular systematic sampling 148
6.3 Estimating mean/total 149
6.4 Estimating mean/total through interpenetrating subsamples 153
6.5 Sample size determination for estimating mean/total 156
6.6 Estimation of proportion 158

6.7 Some further remarks 160
Let us do 161

7. RATIO AND PRODUCT METHODS OF ESTIMATION

7.1 Need for ratio estimation 165
7.2 Estimation of population ratio 166
7.3 Ratio estimator for population mean/total 169
7.4 Determining the sample size for estimation of ratio, mean, and total 175
7.5 Separate and combined ratio estimators 178
7.6 Some further remarks 184
7.7 Product method for estimating mean/total 185
7.8 Determination of sample size for product estimator 189
Let us do 191

8. REGRESSION METHOD OF ESTIMATION

8.1 Introduction 197
8.2 Estimation of mean/total using difference estimator 197
8.3 Estimation of mean/total using estimated regression coefficient 201
8.4 Sample size determination for estimating mean/total 204
8.5 Separate and combined regression estimators 206
8.6 Some further remarks 216
Let us do 217

9. TWO-PHASE SAMPLING

9.1 Need for two-phase sampling 221
9.2 Two-phase sampling in ratio, product, and regression methods of estimation 222
9.3 Sample size determination for ratio, product, and regression estimators 231
9.4 Two-phase PPS sampling 233
9.5 Sampling on two occasions 237
9.6 Some further remarks 242
Let us do 243

10. CLUSTER SAMPLING

10.1 Introduction 248
10.2 Notations 249

10.3 Estimation of mean using simple random sampling 250
10.4 Estimation of total using simple random sampling 258
10.5 Relative efficiency of cluster sampling 264
10.6 Determining the sample size for estimating mean/total 266
10.7 Estimation of proportion 269
10.8 Sample size required for estimation of proportion 273
10.9 Selection of clusters with unequal probabilities 275
10.10 Some further remarks 278
Let us do 278

11. MULTISTAGE SAMPLING

11.1 Introduction 283
11.2 Notations 284
11.3 Estimation of mean/total in two-stage sampling using SRSWOR at both the stages 284
11.4 Estimation of proportion 296
11.5 Estimation of mean/total using PPSWR and SRSWOR 304
11.6 Some further remarks 307
Let us do 308

12. SAMPLING FROM MOBILE POPULATIONS

12.1 Introduction 314
12.2 Estimation of population size using direct sampling 315
12.3 Estimation of population size using inverse sampling 320
12.4 Determining the sample sizes 324
12.5 Some further remarks 328
Let us do 329

13. NONRESPONSE ERRORS

13.1 Introduction 331
13.2 Hansen and Hurwitz technique 332
13.3 Bias reduction without call-backs 337
13.4 Warner's randomized response model 340
13.5 Mangat and Singh's two-stage model 342
13.6 Unrelated question model 345
13.7 Estimation of mean for quantitative characters 350
Let us do 358

APPENDIXES

A.	Standard normal probability distribution	364
B.	Random numbers	365
C.	Number of tractors, tube wells, and net irrigated area (in hectares) for 69 villages of Doraha development block of Punjab, India	369
D.	Fifty WOR simple random samples	371
E.	Explanation of certain local terms used	372

REFERENCES 373

AUTHOR INDEX 383

SUBJECT INDEX 385

Preface

Modern statistics consists of methods which help in drawing inferences about the population under consideration. These populations may actually exist, or could be generated by repeated experimentation. The medium of drawing inferences about the population is the sample, which is a subset of measurements selected from the population. Each measurement in the sample is used for making inferences about the population.

The populations and also the methods of sample selection differ from one field of science to the other. Social scientists use surveys to collect the sample information, whereas the physical scientists employ the method of experimentation for obtaining this information. This is because in social sciences the factors that cause variation in the measurements on the study variable for the population units can not be controlled, whereas in physical sciences these factors can be controlled, at least to some extent, through proper experimental design.

Several excellent books on sampling theory are available in the market. These books discuss the theory of sample surveys in great depth and detail, and are suited to the postgraduate students majoring in statistics. Research workers in the field of sampling methodology can also make use of these books. However, not many suitable books are available, which can be used by the students and researchers in the fields of economics, social sciences, extension education, agriculture, medical sciences, business management, etc. These students and workers usually conduct sample surveys during their research projects. Since they do not have statistics/ mathematics as their major subject, their level of understanding in these fields is not adequate enough. They, therefore, cannot make use of the sampling theory books available in the market, profitably. The present volume is meant to serve these students and research workers.

Besides, the book will also serve as a guide for those professionals, who have to design and conduct surveys as a routine for private and state organizations. The information collected through such surveys may not be used directly for research purposes, but could be used in framing the policies for the welfare of the people.

This book is oriented to discuss various situations that arise in the fields of social sciences, agriculture, business and natural resource management, etc. Suitable methods of sample selection and the estimation procedures for such cases have been considered. This helps in minimizing the survey cost for a given level of estimator's accuracy or in maximizing precision for the budget at disposal.

The book uses only elementary algebraic symbols and will be easily understood by the students and research workers. Each chapter begins with an introduction of the topic, which gives an overview of the content and scope of the chapter besides its linkage with the other chapters. Formulas appropriate to different sampling strategies have been presented without any proofs. It is expected that the readers will try to understand the basic concepts,

and will not be overwhelmed by the formulas. Important definitions and algebraic expressions have been put in boxes so as to enable the reader to have a quick review of the material presented. The pervasive feature that characterizes the present volume is a good number of solved examples. It will help the reader in better understanding of the various steps involved in the calculations of estimator values and their sampling errors from the sample data. Sufficient exercises presenting the applications of methods under discussion are included at the end of each chapter. The reader can use them for further practice. This practice will be of great help in fixing and clarifying different calculation steps involved in any particular case.

The data presented in some of the illustrations and exercises are from actual surveys, while in other cases these may not be real but relate to and reflect real life situations. In certain examples and exercises, the data are given for all the population units. This has been done with a view to clarify certain concepts involved.

There could be readers, who may be interested in having an idea about some other developments that have taken place on the topics considered in the chapter. They may find the section "Some Further Remarks" useful. The references cited, after the appendixes in the book, are intended to arouse readers' curiosity for further learning, and help them in acquiring more detailed accounts of the topics considered.

The material in the book has been arranged into 13 chapters. The first and second chapters are intended to provide the base for subsequent chapters. Chapter 1 discusses briefly the need for sample surveys, and various steps involved in their planning and execution. Important elementary concepts and definitions used in sample surveys are presented in chapter 2. This will help in better understanding of the bulk of the material presented in subsequent chapters. The basic sampling schemes such as simple random sampling, unequal probability sampling, stratified sampling, and systematic sampling are covered in chapters 3 to 6. Chapters 7 and 8 deal with ratio, product, and regression estimators. Other sampling schemes (multiphase, cluster, and multistage sampling) are discussed in chapters 9 through 11. Chapter 12 is devoted to the estimation of size of mobile populations, where a little deviant approach is followed to examine the units. In the last chapter, in addition to the techniques dealing with nonsampling errors, survey methods used for stigmatized and/or sensitive characters are considered . The book in the end also gives author index, subject index, and certain statistical tables. The explanation to certain Indian terms used in the text is given in appendix E. This will facilitate the readers to understand the material presented.

The help we have obtained from the available statistical literature is obvious. The excellent works of W.G. Cochran, M.N. Murthy, P.V. Sukhatme, B.V. Sukhatme, Ms. Shashikala Sukhatme, C. Asok, Des Raj, M.H. Hansen, W.N. Hurwitz, W.G. Madow, F. Yates, S.S. Zarkovich, R.L. Scheaffer, W. Mendenhall, L. Ott, W.E. Deming, L. Kish, I.M. Chakravarti, R.G. Laha, J. Roy, G.A.F. Seber, etc. have been useful in formulating the pattern of development that is followed in the book. All these are gratefully acknowledged. We have also derived some material from the research papers which are listed after the appendixes. Attempts have been made to refer to the original source, but still it is possible that in some instances we might not have recorded proper credit to the authors. We offer them our regrets and apologies.

The American Fisheries Society, Washington, D.C. has our warmest thanks for permission to reprint table 12.1 from their journal *Transactions of the American Fisheries Society.* The permission granted by Indian Statistical Institute, Calcutta, for reproducing the standard normal probability distribution table and the table of random numbers from the book *Formulae and Tables for Statistical Work,* is also gratefully acknowledged .

We record our sincere appreciation to Professor S.K. Mehta for his most perceptive and valuable comments on all the chapters of the book. The authors also express their thanks to S.E.H. Rizvi, M. Bhargava, Inderjit S. Grewal, and M. Javed, doctoral students of the Department of Mathematics and Statistics, for verifying the numerical calculations in the examples and for helping the authors in going through the proofs of the book. The help rendered by Ms. Santosh and Ms. Loveleen in preparing the manuscript is gratefully acknowledged. Thanks are also due to Professor Pritpal Singh Phul, Head, Department of Plant Breeding, for constant encouragement for accomplishment of the project. Many friends, too numerous to mention by name, had made suggestions that are reflected in the final draft of the book.

The authors express their thanks to P.L. Printers, Ludhiana, for preparing an excellent camera-ready version of the manuscript. Ms. Mamta, working with P.L. Printers, has worked hard in typesetting the material in correct and attractive form. Her sincerity of effort is highly appreciated.

We shall be failing in our duty if we do not appreciate the personal sacrifices of our family members: Urmila Singh, Shivendra, Ina, Shailendra, and Shalini; Paramjit Kaur, Parvinder, Gurvinder, and Jaswinder. All these thanks are, however, only a fraction of what is due to the Almighty who granted us an opportunity and mettle to make successful accomplishment of our project.

In spite of the care taken, some errors might have escaped our notice. We shall appreciate if such errors are brought to our attention. Suggestions for improving the present volume will be gratefully received.

Ravindra Singh

Naurang Singh Mangat

CHAPTER 1

Collection of Survey Data

1.1 NEED FOR STATISTICAL DATA

The need to gather information arises in almost every conceivable sphere of human activity. Many of the questions that are subject to common conversation and controversy require numerical data for their resolution. Data resulting from the physical, chemical, and biological experiments in the form of observations are used to test different theories and hypotheses. Various social and economic investigations are carried out through the use and analysis of relevant data. The data collected and analyzed in an objective manner and presented suitably serve as basis for taking policy decisions in different fields of daily life.

The important users of statistical data, among others, include government, industry, business, research institutions, public organizations, and international agencies and organizations. To discharge its various responsibilities, the government needs variety of information regarding different sectors of economy, trade, industrial production, health and mortality, population, livestock, agriculture, forestry, environment, meteorology, and available resources. The inferences drawn from the data help in determining future needs of the nation and also in tackling social and economic problems of people. For instance, the information on cost of living for different categories of people, living in various parts of the country, is of importance in shaping its policies in respect of wages and price levels. Data on health, mortality, and population could be used for formulating policies for checking population growth. Similarly, information on forestry and environment is needed to plan strategies for a cleaner and healthier life. Agricultural production data are of immense use to the state for planning to feed the nation. In case of industry and business, the information is to be collected on labor, cost and quality of production, stock, and demand and supply positions for proper planning of production levels and sales campaigns.

The research institutions need data to verify the earlier findings or to draw new inferences. The data are used by public organizations to assess the state policies, and to point it out to the administration if these are not up to the expectations of the people. The international organizations collect data to present comparative positions of different countries in respect of economy, education, health, culture, etc. Besides, they also use it to frame their policies at the international level for the welfare of people.

1.2 TYPES OF DATA

The collection of required information depends on the nature, object, and scope of study on the one hand and availability of financial resources, time, and man power on the other. The statistical data are of two types: (1) primary data, and (2) secondary data.

Definition 1.1 The data collected by the investigator from the original source are called *primary data.*

Definition 1.2 If the required data had already been collected by some agencies or individuals and are now available in the published or unpublished records, these are known as *secondary data.*

Thus, the primary data when used by some other investigator/agency become secondary data. There could be large number of publications presenting secondary data. Some of the important ones are given below:

1. Official publications of the federal, state, and local governments.
2. Reports of committees and commissions.
3. Publications and reports of business organizations, trade associations, and chambers of commerce.
4. Data released by magazines, journals, and newspapers.
5. Publications of different international organizations like United Nations Organization, World Bank, International Monetary Fund, United Nations Conference on Trade and Development, International Labor Organization, Food and Agricultural Organization, etc.

Caution must be exercised in using secondary data as they may contain errors of transcription from the primary source.

1.3 METHODS OF COLLECTING PRIMARY DATA

There are variety of 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: (1) direct personal interview, (2) questionnaires sent through mail, (3) interview by enumerators, and (4) telephone interview.

1.3.1 *Direct Personal Interview*

In this, the investigator contacts the respondents personally and interviews them. The interviewer asks the questions pertaining to the objective of survey and the information, so obtained, is recorded on a *schedule* (a questionnaire form) already prepared for the purpose. Under this method, the response rate is usually good, and the information is more reliable and correct. However, more expenses and time is required to contact the respondents.

1.3.2 *Questionnaires Sent Through Mail*

In this method, also known as *mail inquiry*, the investigator prepares a questionnaire and sends it by mail to the respondents. The respondents are requested to complete the questionnaires and return them to the investigator by a specified date. The method is suitable where respondents are spread over a wide area. Though the method is less expensive, normally it has a poor response rate. Usually, the response rate in mail surveys has been found to be about 40 percent. The other problem with this method is that it can be adopted only where the respondents are literate and can understand the questions. They

should also be able to send back their responses in writing. The success of this method depends on the skill with which the questionnaire is drafted, and the extent to which willing co-operation of the respondents is secured.

1.3.3 *Interviews by Enumerators*

This method involves the appointment of enumerators by the surveying agency. Enumerators go to the respondents, ask them the questions contained in the schedule, and then fill up the responses in the schedule themselves. For example, the method is used in collecting information during population census. For success of this method, the enumerators should be given proper training for soliciting co-operation of the respondents. The enumerators should be asked to carry with them their identity cards, so that, the respondents are satisfied of their authenticity. They should also be instructed to be patient, polite, and tactful. This method can be usefully employed where the respondents to be covered are illiterate.

1.3.4 *Telephone Interview*

In case the respondents in the population to be covered can be approached by phone, their responses to various questions, included in the schedule, can be obtained over phone. If long distance calls are not involved and only local calls are to be made, this mode of collecting data may also prove quite economical. It is, however, desirable that interviews conducted over the phone are kept short so as to maintain the interest of the respondent.

Payne (1951) and Hyman (1954) have made detailed study of various methods of data collection and associated problems. The books by Murthy (1967) and Des Raj (1968) also contain comments of relevance.

1.4 FRAMING OF QUESTIONNAIRE/SCHEDULE

The *questionnaire* is a channel through which the needed information is elicited. The success of eliciting information, to a considerable extent, depends on the tactful drafting of the questionnaire (or the schedule). The way in which questions are presented affects the quality of response. It is, therefore, important to ensure that not only the right questions are asked but also that they are asked in the right way. This aspect has been dealt with in detail by Murthy (1967) and Des Raj (1968).

Persons framing the questionnaire need to have detailed knowledge of the field of inquiry. While preparing and mailing the questionnaire, following points should be kept in mind:

1. The person conducting the survey must introduce himself and state the objective of the survey. For this purpose, a short letter conveying how the respondent would be benefitted from the survey being conducted, should be enclosed. Also, the enclosing of a self-addressed stamped envelope for the respondent's convenience in returning the questionnaire, will help in improving the response rate.
2. The questions forming the questionnaire/schedule should be clear, unambiguous, and to the point. Vague questions do not bring forth clear and correct answers. As far as possible, questions should be made capable of objective answers. The

language used in the questions should be easy to understand, and the technical terms used should be properly defined.

3. Questions affecting prestige and sentiments of the people and those involving calculations should be avoided while framing the questionnaire. The order of questions should be relevant to generate a logical flow of thought in the minds of respondents. These should not skip back and forth from one topic to another. It will facilitate the answering of each question in turn. It is always advisable to start with simplest questions.
4. Precise and definite instructions for filling the questionnaire and about units of measurement should also be given.
5. The questionnaire should not be lengthy, otherwise, the respondents begin to loose interest in answering them. On the other hand, no important item should be left uncovered.
6. The outlook of questionnaire should be attractive. The printing and the paper used should be of good quality. Sufficient space should be left for answers depending on the type of questions. A pretest of the questionnaire with a group, before its actual use, helps to discover the shortcomings therein. There may be ambiguous questions, the ordering of questions may require change, and some questions may have to be asked in alternate forms. This gives an opportunity to improve the questionnaire in the light of tryout.

1.5 SOME TECHNICAL TERMS

In order to define few technical terms which will be used in the book quite frequently, we consider an example. Let us assume that we wish to find out the proportion of votes a particular political party A, is expected to get in an election in a particular constituency.

Definition 1.3 An *element* is a unit for which information is sought.

In the example considered above, the element will be a registered voter of the constituency. The study variable in this case will be voter's preference for the party A. The variable will be measured as 1 if the voter prefers to vote for party A, otherwise, the measurement will be taken as zero.

Definition 1.4 The *population* or *universe* is an aggregate of elements, about which the inference is to be made.

For the example considered above, population will be the collection of all registered voters of the constituency. It should be noted here that the same population will have different set of measurements for a different study variable.

Populations are called *finite* or *infinite*, depending on the number of units constituting it. The population of registered voters in the above example is finite. Whereas, the populations like water in a tank or a sheet of metal could be considered as infinite populations of their respective molecules. In sample surveys, we shall usually deal with

finite populations. The results for infinite populations could, however, be used for finite populations with very large number of units.

Definition 1.5 *Sampling units* are nonoverlapping collections of elements of the population.

As pointed out earlier, a registered voter is an element in the above example. However, due to convenience or cost considerations, one could sample households in the constituency in place of registered voters, and ask for the preferences of all the registered voters in the sample households. In such a situation, household will be the sampling unit and it may be noted that the number of elements in any sampling unit could be zero, one, or more depending on the number of registered voters in any particular sampled household. If each sampling unit contains one element of the population, then both sampling unit and element are identical.

Definition 1.6 A list of all the units in the population to be sampled is termed *frame* or *sampling frame.*

If individual voter is taken as the sampling unit then a list of all registered voters will constitute the frame. On the other hand, if the households are taken as sampling unit then the list of all households, obtained after properly arranging the list of households in different villages and towns of the constituency, could serve as frame for the selection of a sample of households.

It may be pointed out that the frame may not include all the sampling units of the population at any particular time as the lists of these units are not updated everyday. If frame is the list of registered voters, it may include some voters who have died now, and might not include the names of persons who became eligible to vote after the list of voters was last prepared.

Definition 1.7 A subset of population selected from a frame to draw inferences about a population characteristic is called a *sample.*

In practice number of units selected in a sample is much less than the number in the population. Inferences about the entire population are drawn from the observations made on the study variable for the units selected in the sample. In the example considered above, preference for party A will be asked only from the registered voters selected in the sample. This information will then be used to determine the proportion of all votes that party A is expected to get in the election.

1.6 NEED FOR A SAMPLE

Collection of information on every unit in the population for the characteristics of interest is known as *complete enumeration* or *census*. The money, manpower, and time required for carrying out a census will generally be large, and there are many situations where with limited means complete enumeration is not possible. There are also instances where it is not

feasible to enumerate all units due to their perishable nature. In all such cases, the investigator has no alternative except resorting to a sample survey.

The number of units (not necessarily distinct) included in the sample is known as the *sample size* and is usually denoted by n, whereas the number of units in the population is called *population size* and is denoted by N. The ratio n/N is termed as *sampling fraction.*

There are certain *advantages of a sample survey* over complete enumeration. These are given below :

1.6.1 *Greater Speed*

The time taken for collecting and analyzing the data for a sample is much less than that for a complete enumeration. Often, we come across situations where the information is to be collected within a specified period. In such cases, where time available is short or the population is large, sampling is the only alternative.

1.6.2 *Greater Accuracy*

A census usually involves a huge and unwieldy organization and, therefore, many types of errors may creep in. Sometimes, it may not be possible to control these errors adequately. In sample surveys, the volume of work is considerably reduced. On account of this, the services of better trained and efficient staff can be obtained without much difficulty. This will help in producing more accurate results than those for complete enumeration.

1.6.3 *More Detailed Information*

As the number of units in a sample are much less than those in census, it is, therefore, possible to observe/interview each and every sample unit intensively. Also, the information can be obtained on more number of variables. However, in complete enumeration such an effort becomes comparatively difficult.

1.6.4 *Reduced Cost*

Because of lesser number of units in the sample in comparison to the population, considerable time, money, and energy are saved in observing the sample units in relation to the situation where all units in the population are to be covered.

From the above discussion, it is seen that the sample survey is more economical, provides more accurate information, and has greater scope in subject coverage as compared to a complete enumeration. It may, however, be pointed out here that *sampling errors* are present in the results of the sample surveys. This is due to the fact that only a part of the whole population is surveyed. On the other hand, *nonsampling errors* are likely to be more in case of a census study than these are in a sample survey. Merits and demerits of sample surveys have been discussed in detail by Zarkovich (1961) and Lahiri (1963).

1.7 SAMPLING PROCEDURES

The method which is used to select the sample from a population is known as *sampling procedure.* These procedures can be put into two categories - probability sampling and nonprobability sampling. These two types of surveys are not distinguished by the questionnaire and instructions to be followed, but by the methods of selecting the sample for obtaining the estimates of the population characteristics of interest and their precision.

1.7.1 *Probability Sampling*

Definition 1.8 If the units in the sample are selected using some probability mechanism, such a procedure is called *probability sampling.*

This type of survey assigns to each unit in the population a definite probability of being selected in the sample. Alternatively, it enables us to define a set of distinct samples which the procedure is capable of selecting if applied to a specific population. The sampling procedure assigns to each possible sample a known probability of being selected. One can build suitable estimators for different population characteristics for probability samples. For any sampling procedure of this type, one is in a position to develop theory by using probability apparatus. It is also possible to obtain frequency distribution of the estimator values it generates if repeatedly applied to the same population. The measure of the sampling variation can also be obtained for such procedures, and the proportion of estimates that will fall in a specified interval around the true value can be worked out. The procedures such as these will only be considered in this book.

1.7.2 *Nonprobability Sampling*

Definition 1.9 The procedure of selecting a sample without using any probability mechanism is termed as *nonprobability sampling.*

The convenience sampling and the purposive sampling belong to this category. In *convenience sampling,* the sample is restricted to a part of the population that is readily accessible. For example, a sample of coal from an open wagon may be taken from the depth of up to 50 cm from the top. In studies where the process of taking observations is inconvenient, unpleasant, or troublesome to the selected person, only the volunteers may constitute the sample.

Purposive sampling (also termed *Judgement sampling*) is common when special skills are required to form a representative subset of population. For instance, auditors often use judgement samples to select items for study to determine whether a complete audit of items may be necessary. Sometimes, quotas are fixed for different categories of population based on considerations relevant to the study being conducted, and selections within the categories are based on personal judgement. This type of sampling procedure is also termed *quota sampling.*

Obviously, these methods are subject to human bias. In appropriate conditions, these methods can provide useful results. They are, however, not amenable to the development of relevant theory and statistical analysis. In such methods, the sampling error can not be objectively determined. Hence, they are not comparable with the available probability sampling methods.

1.8 WITH AND WITHOUT REPLACEMENT SAMPLING

Definition 1.10 In *with replacement* (WR) *sampling,* the units are drawn one by one from the population, replacing the unit selected at any particular draw before executing the next draw.

As the constitution of population remains same at each draw, some units in the with replacement sample may get selected more than once. This procedure gives rise to N^n possible samples when order of selection of units in the sample is taken into account, where N and n denote the population and sample sizes respectively.

Example 1.1

Given below are the weights (in pounds) of 4 children at the time of birth in a hospital:

Child :	A	B	C	D
Weight :	5.5	8.0	6.5	7.0

Enumerate all possible WR samples of size 2. Also, write values of the study variable (weight) for the sample units.

Solution

Here, N=4 and n=2. There will, therefore, be $4^2=16$ possible samples. These are enumerated below along with the weight values for the units included in the sample.

Table 1.1 Possible samples along with their variable values

Sample	Children in the sample	Weight for the sampled children	Sample	Children in the sample	Weight for the sampled children
1	A, A	5.5, 5.5	9	C, A	6.5, 5.5
2	A, B	5.5, 8.0	10	C, B	6.5, 8.0
3	A, C	5.5, 6.5	11	C, C	6.5, 6.5
4	A, D	5.5, 7.0	12	C, D	6.5, 7.0
5	B, A	8.0, 5.5	13	D, A	7.0, 5.5
6	B, B	8.0, 8.0	14	D, B	7.0, 8.0
7	B, C	8.0, 6.5	15	D, C	7.0, 6.5
8	B, D	8.0, 7.0	16	D, D	7.0, 7.0

■

Definition 1.11 In *without replacement* (WOR) *sampling*, the units are selected one by one from the population, and the unit selected at any particular draw is not replaced back to the population before selecting a unit at the next draw.

Obviously, no unit is selected more than once in a WOR sample. If the order of selection of units in the sample is ignored, then there are $\binom{N}{n}$ possible samples for this selection procedure.

Example 1.2

Using data of example 1.1, enumerate all possible WOR samples of size 2, and also list the weight values for the respective sample units.

Solution

In this case, number of possible samples will be $\binom{4}{2} = 6$. These are enumerated below. Note that no samples like AA or BB appear in the list of possible samples, and also the ordered samples like AB and BA are treated as the same sample.

Sample	Children in the sample	Weight for the sample children	Sample	Children in the sample	Weight for the sample children
1	A, B	5.5, 8.0	4	B, C	8.0, 6.5
2	A, C	5.5, 6.5	5	B, D	8.0, 7.0
3	A, D	5.5, 7.0	6	C, D	6.5, 7.0 ■

1.9 PLANNING AND EXECUTION OF SAMPLE SURVEYS

Sample survey techniques are used widely as an organized and fact finding instrument. The quality of the inferences drawn about the population characteristics from the sample data is related to, how well, the sample represents the population. It requires to select a suitable sampling plan, and implement it in a way that ensures the sample to be a good representative of the population under study. It is, therefore, essential to describe briefly the steps involved in the *planning* and *execution* of a survey. Surveys vary greatly in their scope and complexity. Problems that are baffling in one survey, may be trivial or nonexistent in another. Some of the important aspects requiring attention at the planning stage are grouped under the following heads:

1.9.1 *Objectives*

The first task is to lay down, in concrete terms, the objectives of the survey. The investigator should ensure that these *objectives* are commensurate with available resources in terms of money, manpower, and the time limit specified for the survey.

1.9.2 *Population to be Studied*

The *population* to be covered by the survey should be clearly defined. An exact description should be given of the geographical region and the categories of the material to be covered by the survey. For instance, in a survey of human population, it is necessary to specify whether such categories as hotel residents, institutions, military personnel, etc., were to be included or not.

Population to be sampled should coincide with the *target population* about which inferences are to be drawn. However, sometimes impracticability and inconvenience may result in the leaving out of certain segments of the population from the scope of the survey. If so, the conclusions drawn will apply only to the *population* actually *sampled.* Any supplementary information gathered for the omitted sectors, which can throw some light on the subject matter of the survey, will be useful.

1.9.3 *Sampling Unit*

The population should be capable of being divided into *sampling units*, and these should be properly defined. For example, a human population can be considered to be built up of villages, localities, households, persons, etc. The division of population into sampling units should be unambiguous. Every element of the population should correspond to just one and only one sampling unit. The border line cases can be handled by framing some appropriate rules.

1.9.4 *The Sampling Frame*

In surveys, as already discussed, it is always desired that the sampled and the target population should coincide. It should, therefore, be ensured that all the sampling units of the population under study are included in the frame. The frame should be up to date and free from errors of omission and overlapping.

1.9.5 *Sample Selection*

The size of the sample and manner of selecting the sample should receive careful attention. After taking various technical, operational, and risk factors into consideration, an optimum size of the sample and sampling procedure need to be decided upon. While doing so, the aim of achieving either a given degree of precision with a minimum cost, or the maximum precision with a fixed cost, should be kept in mind. It should also be ensured that the sample is representative of the population.

1.9.6 *Methods of Collecting Information*

After a careful examination of the frame, the method of sample selection, available resources, and the objectives of survey, one should decide about the type of data to be used, that means, whether to collect primary data or to use secondary data. In case the primary data are to be collected, the investigator should decide whether data are to be collected by personal face-to-face interview, by mail, through enumerators, or by telephone interview. These methods have already been briefly discussed in section 1.3.

1.9.7 *Handling of Nonresponse*

Procedures should be devised to deal with the respondents, who do not give information by choice, or are not found at home . The reason for nonresponse should also be ascertained. This helps in assessing the effect of refusals and random nonresponse on the conclusions to be drawn.

1.9.8 *Pilot Survey*

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 past surveys, it is desirable to design and carry out a *pilot survey*. It will be useful for : (1) discovering shortcomings in the questionnaire/ schedule, (2) evolving suitable strategies for field and analysis work, and (3) training the staff for the purpose.

1.9.9 *Organization of Field Work*

Different aspects of *field work* such as recruitment and training of investigators, and inspection and supervision of field staff should be given due consideration in the light of the prevailing operational conditions. The personnel engaged in the survey must receive training, not only in the purpose of the survey and in the methods of measurement to be employed, but also in the art of eliciting acceptable responses. The investigators should be able to withstand long and arduous travel, sometimes in inhospitable conditions. The work must be adequately supervised, as it is important for the investigator to adhere to procedures and tact in answering the questions raised by respondents. Besides, it will help in resolving unusual or unforeseen problems in the field.

A quality check and editing need to be instituted to make careful review of questionnaires received. It will be valuable in amending the recording errors and deleting the data that are erroneous and superfluous.

1.9.10 *Analysis of Data and Preparation of Report*

The stage of analysis of collected data and drawing inferences from a sample is a vital issue, as the results of survey are the backbone of the policies to be framed. The errors creeping in the tabulation and statistical analysis of data should be kept under control.

Last, but not the least, comes the *report writing*. While writing the report, the objectives, the scope, and the subject coverage must be mentioned. It is also essential to clarify the method of data collection, estimation procedure including tabulation and analysis, and cost structure in the report. A brief description of the organizations sponsoring and conducting the survey should also be included. Relevant published papers and reports should be cited for reference. The report should conclude with a summary of findings and suggestions for possible action to be taken. For a report on an actual sample survey, the reader may refer to Des Raj (1968).

Many interesting examples showing the range of applications of the sampling methods in business have been given by Deming (1960) and Slonim (1960).

LET US DO

1.1 Discuss the statement: "The need to collect statistical information arises in almost every conceivable sphere of human activity."

1.2 Describe briefly each of the following terms:

a. Primary data	b. Secondary data
c. Mail inquiry	d. Questionnaire/schedule
e. Population	f. Census
g. Element	h. Sample
i. Sampling unit	j. Sampling frame

1.3 Distinguish between primary and secondary data. Give examples. Which of the two data are more reliable and why ?

1.4 What precautions should one take in making use of published data for further studies ?

1.5 Discuss the methods usually employed in collection of primary data, and state briefly their respective merits and demerits.

1.6 In the situations given below, define universe and indicate which method of data collection - personal interview, mail inquiry, or telephone interview - would you prefer, keeping the cost involved and other relevant factors in view:

a. The investigator is interested in estimating the percent loss caused to wheat crop by hail storm in 500 villages. The elected head of the village (known as *sarpanch*) is the sampling unit.
b. To estimate consumer acceptance of the newly developed car model, before it is introduced in the market.
c. The investigator wishes to estimate the average time taken by a telephone company to attend to the complaints of its customers.
d. A firm wishes to estimate the gas mileage for its newly developed model of car. The addresses of the buyers of all the cars of this model are available in the head office of the firm.

1.7 Differentiate between target and sampled population. What problem arises if two populations are not same ?

1.8 What is a questionnaire/schedule? What are the various considerations one should keep in mind while framing it ?

1.9 Distinguish between a questionnaire and a schedule. What is each used for ?

1.10 A publisher wishes to determine, whether his writers prefer title of their books in bright or dull colors? How would you define population, and what sort of questionnaire would you draft to get an answer to the problem?

1.11 The authorities of a certain university wish to conduct a survey to elicit views of its faculty about introducing five days week in the university. Suggest possible sampling unit and frame for the purpose. Also, draft a suitable questionnaire for use in this survey.

1.12 Discuss relative merits of using a sample survey in relation to a census.

1.13 Distinguish between a sample inquiry and complete enumeration. Under what circumstances can the latter be recommended in preference to sample survey ?

1.14 In the following situations, indicate whether a sample study or a complete enumeration should be undertaken, and why ?

a. A firm wishes to estimate the longevity of 800 electric bulbs manufactured of a new design.
b. A tractor manufacturing company is interested in obtaining information on customer preferences in respect of parasol that protects the driver from the sun and rain.
c. A university has 600 teachers. The administration wishes to determine the acceptability of subscribing to a new group insurance scheme.

1.15 Do you agree that it is possible for sample results to be more accurate than the census results ? If so, explain.

1.16 What is the primary advantage of probability sampling over the nonprobability sampling ? Cite three situations where nonprobability sampling is to be preferred.

1.17 Describe briefly the difference between with and without replacement sampling.

1.18 Consider a population consisting of 6 villages, the areas (in hectares) of which are given below :

Village :	A	B	C	D	E	F
Area :	760	343	657	550	480	935

a. Enumerate all possible WR samples of size 3. Also, write the values of the study variable for the sampled units.

b. List all the WOR samples of size 4 along with their area values.

1.19 The Department of Agriculture desires to estimate yield per hectare of wheat in a district. Describe the various steps that may be involved in the planning and execution of a sample survey for this purpose.

CHAPTER 2

Elementary Concepts

2.1 INTRODUCTION

Knowledge of basic concepts is a prerequisite for an insight into the sample survey designs. Assuming some exposure to elementary probability theory on the part of the reader, we present in this chapter, a rapid review of some of these concepts. To begin with, preliminary statistical concepts including those of expectation, variance, and covariance, for random variables and linear functions of random variables will be defined. The idea of sampling distribution, being basic to the sampling theory, has been briefly explained. The concepts of measure of error, interval estimation, and sample size determination, which are related to sampling distribution, have also been discussed. The chapter concludes with a brief introduction to the sampling and nonsampling errors.

The *notations* to be used in the book will be defined, as far as possible, wherever they first appear. Usually, the uppercase letters will be used to denote the values of variables for population units, whereas lowercase letters will denote the corresponding values for sample units. The population parameters are denoted either by the uppercase letters of English alphabet or by Greek letters. The respective point estimators of these parameters will be denoted either by lowercase English letters, or by putting caps (^) on the corresponding symbols of parameters.

2.2 STATISTICAL PRELIMINARIES

Assuming some knowledge of the probability theory on the part of the reader, we briefly review here the concepts of expectation, variance, and covariance.

2.2.1 *Expectation*

Definition 2.1 If x is a random variable which assumes values $x_1, x_2, ..., x_k$ with probabilities $p_1, p_2, ..., p_k$ respectively with $\Sigma\ p_i = 1$, i= 1,2, ..., k, then *expectation* of the variable x is defined as

$$E(x) = \sum_{i=1}^{k} p_i x_i \qquad (2.1)$$

In physical sense, expected value of the random variable x stands for the center of gravity of the probability distribution, where a probability mass p_i is located at the point $x=x_i$, i=1, 2, ..., k.

Example 2.1
A fair die is rolled once. If x denotes the number on the upper face of the die, find expected value of x.

Solution
Here, the random variable x is the number on the upper face of the die. Thus, x can take any one of the values 1, 2,...,6, each with probability 1/6. Hence using (2.1),

$$E(x) = \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6)$$

$$= 3.5 \blacksquare$$

We now state some more useful related results.

Result 2.1 Let a and b be two constants and x a random variable, then

$$E(ax+b) = aE(x)+b$$

Result 2.2 If $x_1, x_2,...,x_k$ are random variables, and $a_1, a_2,..., a_k$ are constants, then

$$E(a_1x_1+a_2x_2+...+a_kx_k) = a_1E(x_1)+a_2E(x_2)+...+a_kE(x_k)$$

Result 2.3 If $x_1,x_2,...,x_k$ are k mutually independent random variables, then

$$E(x_1.x_2...x_k) = E(x_1).E(x_2)...E(x_k)$$

Result 2.4 Let x and y be two random variables which are not independent, then

$$E(xy) = E_1[xE_2(y)]$$

where E_2 is the conditional expectation of y for a given value of x, and E_1, the expectation over all values of x.

2.2.2 *Variance and Covariance*

If x is a random variable and E(x) its expected value, then variance of x is defined as

$$\left.\begin{aligned} V(x) &= E[x-E(x)]^2 \\ &= E(x^2)-[E(x)]^2 \end{aligned}\right] \qquad (2.2)$$

When the investigator is interested in linear dependence of pairs of random variables, such as income x and expenditure y, then it is indicated by a measure called *covariance* of x and y. This term is defined as

$$Cov(x,y) = E[\{x-E(x)\}\{y-E(y)\}]$$

A zero value of the covariance indicates no linear dependence between x and y.

Example 2.2
For the experiment given in example 2.1, determine the variance of random variable x.

Solution
First we work out the term $E(x^2)$. This will be

$$E(x^2) = \frac{1}{6}(1)^2 + \frac{1}{6}(2)^2 + \frac{1}{6}(3)^2 + \frac{1}{6}(4)^2 + \frac{1}{6}(5)^2 + \frac{1}{6}(6)^2$$
$$= \frac{91}{6}$$
$$= 15.167$$

On substituting in (2.2) the value of $E(x^2)$ obtained above, and of $E(x)$ from example 2.1, one gets

$$V(x) = 15.167 - (3.5)^2$$
$$= 2.917 \blacksquare$$

2.2.3 *Variance of Linear Functions of Random Variables*

In the theory of sample surveys, the investigator often requires the variance of a linear function of random variables to determine the amount of error in the estimator. Following result will be helpful in such a case.

Result 2.5 Let $x_1, x_2,...,x_k$ be k random variables, then

$$V\left(\sum_{i=1}^{k} a_i x_i\right) = \sum_{i=1}^{k} a_i^2 V(x_i) + 2\sum\sum_{i<j} a_i a_j \text{Cov}(x_i, x_j)$$

where $a_1, a_2,..., a_k$ are constants. The double summation taken over all pairs (x_i, x_j), with $i < j$, will involve $k(k-1)/2$ covariance terms.

For independence of random variables x_i and x_j, $\text{Cov}(x_i, x_j)=0$. Thus, result 2.5 gets simplified in case of mutually independent random variables as

$$V\left(\sum_{i=1}^{k} a_i x_i\right) = \sum_{i=1}^{k} a_i^2 V(x_i)$$

The above expression holds even when the random variables $x_1, x_2,..., x_k$, are mutually uncorrelated. The reader should note, that the condition of no correlation is much less stringent than the condition of independence.

2.3 ESTIMATOR AND ITS SAMPLING DISTRIBUTION

The ultimate objective of any sample survey is to make inferences about a population of interest. Such inferences are based on information contained in a sample selected from that population. The investigator usually aims at the estimation of certain unknown features of the population. These population characteristics are called parameters.

Definition 2.2 Any real valued function of variable values for all the population units is known as a *population parameter* or simply a *parameter.*

For any given variable, the population value of a parameter is constant. Some of the important parameters frequently required to be estimated in surveys are total, mean, proportion, and variance. For instance, if $Y_1, Y_2, ..., Y_N$ are the values of the variable y for the N units in the population, then

$$\text{Population mean} = \bar{Y} = \frac{Y_1 + Y_2 + ... + Y_N}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i \tag{2.3}$$

$$\text{Population variance} = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right) \tag{2.4}$$

Definition 2.3 A real valued function of variable values for the units in the sample is called a *statistic.* If it is used to estimate a parameter, it is termed as *estimator.*

The particular value taken by the estimator for a given sample, is known as *estimate* or *point estimate.* For instance, the mean for a given sample, provides an estimate of population mean. The *sample mean* $\bar{y}$ and *sample mean square* s^2 for a sample of size n, are respectively given by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.5}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) \tag{2.6}$$

where $y_1, y_2, ..., y_n$ are the values of the variable y for the n units in the sample. Note that the *standard deviation* is the positive square root of the variance. It will, therefore, be denoted by σ for the population.

The probability mechanism underlying the process of sample selection usually gives rise to different samples. The estimates based on sample observations may differ from sample to sample, and also from the true value of the parameter. The estimator, therefore, is a *random variable.* It leads us to the concept of sampling distribution.

Definition 2.4 For a given population, sampling procedure, and sample size, the array of possible values of an estimator each with its probability of occurrence, is the *sampling distribution* of that estimator.

Example 2.3

Four cows in a household marked A, B, C, and D respectively yield 5.00, 5.50, 6.00, and 6.50 kg of milk per day. Obtain the sampling distribution of average milk yield based on samples of n=2 cows, when the cows are selected with equal probabilities and WR. The procedure for drawing a sample has been explained in chapter 3.

Solution

Here N=4 and n=2. The number of possible samples in this case will be $4^2=16$. The mean of each sample, along with the cows included in the sample is given in table 2.1.

Table 2.1 Means of all possible samples

Sample	Cows in the sample	Sample mean $\bar{y}$	Sample	Cows in the sample	Sample mean $\bar{y}$
1	A, A	5.00	9	C, A	5.50
2	A, B	5.25	10	C, B	5.75
3	A, C	5.50	11	C, C	6.00
4	A, D	5.75	12	C, D	6.25
5	B, A	5.25	13	D, A	5.75
6	B, B	5.50	14	D, B	6.00
7	B, C	5.75	15	D, C	6.25
8	B, D	6.00	16	D, D	6.50

From the above table, we get the following sampling distribution :

Table 2.2 Distribution of sample mean

Sample mean ($\bar{y}$)	Frequency (f)	Probability (p)
5.00	1	.0625
5.25	2	.1250
5.50	3	.1875
5.75	4	.2500
6.00	3	.1875
6.25	2	.1250
6.50	1	.0625
Total	16	1

The sampling distribution for any estimator can be used for finding out probabilities for various statements about the values taken by the estimator. In this particular case, one can easily verify that the probability P ($\bar{y} \geq 6$) = 6/16. Similarly, P ($\bar{y} < 5.25$) =1/16, or P ($5.25 \leq \bar{y} \leq 5.75$) = 9/16. ■

In this case, we had considered a population of just four cows (units). However, in practice the number of units in the population will generally be quite large, and so will be the number of units to be selected in the sample. This will make the number of possible samples, and consequently the number of values taken by the sample mean (estimator), very large. In such cases, the enumeration of all possible samples, and also of the values taken by the estimator, will be quite difficult. One will, therefore, not be in a position to easily determine the exact sampling distribution of an estimator.

Also in the example considered above, all possible samples had equal chance of being selected. In practice, however, there may be situations where different samples will have different probabilities of selection. To find out exact probabilities for statements regarding the estimator values in such cases, one will be required to add the probabilities of selection of all those samples which yield estimator values satisfying the statement. This may not be an easy job to do in most of the real life situations. Exact sampling distribution of estimators in all such cases may, therefore, have to be approximated by some known continuous probability density functions to make the calculation of probabilities for various statements about the estimator value, easier.

For continuous random variables, if the study variable is normally distributed in the population, then the distribution of sample mean is exactly normal for any sample size. However, if the study variable is not normally distributed in the population, the distribution of sample mean approaches *normal distribution* as the sample size n is increased.

2.4 UNBIASED ESTIMATOR

In survey sampling, a good estimator is expected to have mainly two properties. One of these properties is known as unbiasedness. The other one, which we consider in the next section, is the closeness of the values taken by the estimator for different possible samples, to the actual unknown value of the parameter.

For discussion in the remaining part of this chapter, we shall denote the population parameter in general by θ, whereas its estimator will be denoted by $\hat{\theta}$. In the preceding section, we have observed that the estimator $\hat{\theta}$ is a random variable, and it takes different values for different possible samples that can be selected from the population under consideration. The sample selection procedure generates a sampling distribution for $\hat{\theta}$. The unbiasedness of the estimator $\hat{\theta}$ is then defined as follows :

Definition 2.5 The estimator $\hat{\theta}$ is said to be *unbiased* for the parameter θ, if $E(\hat{\theta}) = \theta$.

If the sample selection procedure is such that all possible samples are equally likely to get selected, $E(\hat{\theta})$ becomes the simple average of the values that the estimator $\hat{\theta}$ takes

for different possible samples which can be selected from the population under study. However, if the probabilities of selection for the different possible samples are not equal, then E ($\hat{\theta}$) will be the weighted average (weights being the probabilities of selection for different possible samples) of the values of $\hat{\theta}$. Thus unbiasedness of an estimator $\hat{\theta}$ ensures, that on the average it will take value equal to the unknown population parameter θ, although for most of the samples, the values taken by $\hat{\theta}$ will be either less or more than θ.

Example 2.4

For the data in example 2.3, verify whether the sample mean based on n=2 cows is an unbiased estimator for the average milk yield in the population ? Assume that the cows in the sample are selected with equal probabilities and with replacement.

Solution

In table 2.1 are given the sample mean values for 16 possible samples of size n=2 that can be selected from a population of size N=4 with equal probabilities and with replacement. Here, sampling procedure ensures that all these samples have same chance of being selected. Hence, expected value of sample mean $\bar{y}$ will be the simple average of 16 possible sample mean values. Thus,

$$E(\bar{y}) = \frac{1}{16}(5.00+5.25+...+6.50)$$

$$= \frac{92.00}{16}$$

$$= 5.75$$

Using (2.3), it can be easily seen that the population mean $\bar{Y}$ (the parameter θ in this case) is also equal to 5.75. Hence, the sample mean $\bar{y}$ is unbiased for the population mean $\bar{Y}$. ■

In case where E ($\hat{\theta}$) is not equal to the value of population parameter θ, the estimator $\hat{\theta}$ is said to be a biased estimator.

Definition 2.6 If for an estimator $\hat{\theta}$, $E(\hat{\theta}) \neq \theta$,the estimator $\hat{\theta}$ is called a *biased estimator* of θ. The magnitude of the bias in $\hat{\theta}$ is given by

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

The ratio

$$RB(\hat{\theta}) = \frac{B(\hat{\theta})}{\theta}$$

is called the *relative bias* of the estimator $\hat{\theta}$.

2.5 MEASURES OF ERROR

As discussed in section 2.4, it is not sufficient that an estimator be unbiased for it to qualify as a good estimator. In addition to the property of unbiasedness, the estimator should also have small sampling variance. As pointed out earlier, the value of the estimator $\hat{\theta}$ may differ from sample to sample and also from its parameter value θ. The differences $(\hat{\theta}_s - \theta)$ and $[\hat{\theta}_s - E(\hat{\theta})]$, $\hat{\theta}_s$ being the estimator's value based on the s-th sample, denote the error of the estimate $\hat{\theta}_s$. On pooling such errors for all possible samples that can be selected from the population under consideration, one gets a single measure of error for the sampling situation at hand. Some commonly used measures of this error are presented here.

2.5.1 *Sampling Variance*

Definition 2.7 The *sampling variance* is a measure of the divergence of the estimator values from its expected value. Alternatively, it is the variance of the sampling distribution of an estimator. In the light of (2.1) and (2.2), one gets it as

$$V(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$$
$$= E(\hat{\theta})^2 - [E(\hat{\theta})]^2$$

The positive square root of sampling variance is termed *standard error* (SE). Thus,

$$SE(\hat{\theta}) = +\sqrt{V(\hat{\theta})}$$

In other words, SE is the standard deviation of the sampling distribution. It is also an important measure of the fluctuations in the estimator values due to specific sampling design.

Example 2.5
For the data in example 2.3, compute the sampling variance and standard error of sample mean for the WR equal probability samples of size 2.

Solution
The sampling distribution of mean $\bar{y}$ (here the estimator $\hat{\theta}$ is the sample mean $\bar{y}$) for n=2 has been obtained in example 2.3. Also from example 2.4, $E(\bar{y}) = 5.75$. Thus from table 2.1, we have

$$V(\bar{y}) = \frac{1}{16}[(5.00)^2 + (5.25)^2 + \ldots + (6.50)^2] - (5.75)^2$$

which from table 2.2, is equivalent to

$$V(\bar{y}) = \frac{1}{16}[(5.00)^2(1) + (5.25)^2(2) + ... + (6.50)^2(1)] - (5.75)^2$$
$$= .15625$$

It gives

$$SE(\bar{y}) = \sqrt{.15625}$$
$$= .39528. \blacksquare$$

It should be noted that the sampling variance $V(\hat{\theta})$ and $SE(\hat{\theta})$ are still of no practical use, because their values depend on the study variable values for all the population units which are usually not available in practice. Thus to get an idea about the magnitude of the error involved in $\hat{\theta}$ values, one needs to estimate $V(\hat{\theta})$ and $SE(\hat{\theta})$ from the sample data. Their estimators are respectively denoted by $v(\hat{\theta})$ and $se(\hat{\theta})$. The term $se(\hat{\theta})$, called the *estimate of standard error* of estimator $\hat{\theta}$, is the positive square root of $v(\hat{\theta})$. Thus,

$$se(\hat{\theta}) = + \sqrt{v(\hat{\theta})}$$

2.5.2 *Mean Square Error*

In case the estimator is biased, we use mean square error for measuring the variability of sampling distribution.

Definition 2.8 The *mean square error* (MSE) measures the divergence of the estimator values from the true parameter value. This can be put as

$$MSE(\hat{\theta}) = E(\hat{\theta}-\theta)^2$$

The positive square root of MSE is termed as *root mean square error*. The MSE and the sampling variance are related as

$$MSE(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2$$

where $B(\hat{\theta})$ is the bias of the estimator $\hat{\theta}$. Thus, for an unbiased estimator , the $MSE(\hat{\theta})$ and the $V(\hat{\theta})$ are equivalent. The above discussion about MSE now enables us to talk about relative efficiency.

Definition 2.9 If $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of the parameter θ, the *relative efficiency* of the estimator $\hat{\theta}_2$ with respect to the estimator $\hat{\theta}_1$, is defined as

$$RE = \frac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)} \qquad (2.7)$$

Thus for the estimator $\hat{\theta}_2$ to be more efficient than the estimator $\hat{\theta}_1$, the RE defined above will be more than one. Since the mean square errors, MSE($\hat{\theta}_1$) and MSE($\hat{\theta}_2$), are not known in practice, their respective sample estimates denoted by mse($\hat{\theta}_1$) and mse($\hat{\theta}_2$) are used in their place.

Example 2.6

The estimated mean square errors of two estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, are 4861.79 and 5258.62 respectively. Estimate percent relative efficiency of estimator $\hat{\theta}_2$ with respect to $\hat{\theta}_1$. Also point out, which of the two estimators is more efficient ?

Solution

We have mse($\hat{\theta}_1$)=4861.79 and mse($\hat{\theta}_2$)=5258.62. Through relation (2.7), the estimate of RE of estimator $\hat{\theta}_2$ in relation to $\hat{\theta}_1$ would be

$$RE = \frac{\text{mse}\,(\hat{\theta}_1)}{\text{mse}\,(\hat{\theta}_2)}$$

On making substitutions, one gets

$$RE = \frac{4861.79}{5258.62}$$

$$= .925$$

Percent RE = .925 (100)

= 92.5

As the RE obtained is less than 1, the estimator $\hat{\theta}_2$ is less efficient as compared to the estimator $\hat{\theta}_1$. ■

In addition to the variance and the mean square error, there are two more terms that are sometimes used in literature. These are *accuracy* and *precision* which refer to the deviations of $\hat{\theta}$ values from θ and $E(\hat{\theta})$ respectively.

Remark 2.1 Another property of a good estimator, besides unbiasedness and small mean square error, is *consistency*. An estimator $\hat{\theta}$ is said to be a *consistent estimator* of parameter θ, if it approaches θ with probability tending to unity as the sample size tends to infinity. This definition of consistency, thus, strictly applies to estimators based on samples drawn from infinite populations. In case of finite populations, the estimator $\hat{\theta}$ is said to be a consistent estimator of θ if it assumes value θ when the entire population is taken as the sample. An easy way to find out whether any particular estimator $\hat{\theta}$ is a consistent estimator or not, is to find the limit of MSE ($\hat{\theta}$) as the sample size tends to infinity. If this limit is zero, then the estimator is consistent.

2.6 CONFIDENCE INTERVALS

The point estimates yielded by different samples are rarely equal to the parameter value. But we can enhance the usefulness of estimation by calculating a specific region, known

as *confidence interval*, in which the population parameter probably lies. We may also be able to give probability of this confidence interval covering the parameter. The confidence interval (also called *interval estimate*) of a parameter θ is an interval of the form $\hat{\theta}_l \le \theta \le \hat{\theta}_u$. The values of $\hat{\theta}_l$ and $\hat{\theta}_u$, the *lower* and *upper confidence limits* of the interval, depend on the value taken by the estimator $\hat{\theta}$ for a given sample, and also on the sampling distribution of $\hat{\theta}$ (or of a suitable function $h(\hat{\theta})$ of $\hat{\theta}$ for which the sampling distribution is already known). Based on the known sampling distribution of the function $h(\hat{\theta})$, we choose $h_l(\hat{\theta})$ and $h_u(\hat{\theta})$, such that for any specified probability $(1-\alpha)$, where $0 < \alpha < 1$, $P[h(\hat{\theta}) < h_l(\hat{\theta})] = P[h(\hat{\theta}) > h_u(\hat{\theta})] = \alpha/2$ and $P[h_l(\hat{\theta}) \le h(\hat{\theta}) \le h_u(\hat{\theta})] = 1-\alpha$. Knowing the form of the function $h(\hat{\theta})$, the double inequality $h_l(\hat{\theta}) \le h(\hat{\theta}) \le h_u(\hat{\theta})$ can, through algebraic manipulation, be put as $\hat{\theta}_l \le \theta \le \hat{\theta}_u$. Thus, whenever the double inequality $h_l(\hat{\theta}) \le h(\hat{\theta}) \le h_u(\hat{\theta})$ holds, the inequality $\hat{\theta}_l \le \theta \le \hat{\theta}_u$ also holds. Hence, the probability $P(\hat{\theta}_l \le \theta \le \hat{\theta}_u)$ is also equal to $(1-\alpha)$. Such an interval $\hat{\theta}_l \le \theta \le \hat{\theta}_u$, computed from a particular sample, is known as $(1-\alpha)100$ percent confidence interval for θ. The probability $(1-\alpha)$ is called the *confidence coefficient* for the interval.

To illustrate, let us assume that the estimator $\hat{\theta}$ is normally distributed with mean θ and known variance $V(\hat{\theta})$. The variance of the estimator $\hat{\theta}$ may be known from some previous survey, or from a pilot study. The simple function $Z=(\hat{\theta}-\theta)/\sqrt{V(\hat{\theta})}$ of $\hat{\theta}$,will then be following standard normal distribution which is extensively tabulated. An abridged version of the tables of probabilities, for this distributon, are given in appendix A. From the standard normal probability tables, the value of $Z_{\alpha/2}$ that satisfies the relation $P(Z<-Z_{\alpha/2}) = P(Z>Z_{\alpha/2}) = \alpha/2$, can be easily determined for any value of α. If $\alpha=.05$, it is then easily seen that $Z_{\alpha/2}=Z_{.025}=1.96$, since $P(Z < -1.96) = P(Z > 1.96) = .025$. Thus, $P[-1.96 \le Z= (\hat{\theta}-\theta)/\sqrt{V(\hat{\theta})} \le 1.96]=1-.05 =.95$. This probability statement is equivalent to $P[\hat{\theta}-1.96\sqrt{V(\hat{\theta})} \le \theta \le \hat{\theta}+1.96\sqrt{V(\hat{\theta})}]=.95$. Therefore,$\hat{\theta}_l = \hat{\theta}-1.96\sqrt{V(\hat{\theta})}$ and $\hat{\theta}_u = \hat{\theta}+1.96\sqrt{V(\hat{\theta})}$, and the confidence interval $[\hat{\theta}-1.96\sqrt{V(\hat{\theta})}, \hat{\theta}+1.96\sqrt{V(\hat{\theta})}]$ has the confidence coefficient of .95. That means, it will cover the unknown population parameter θ with probability .95. Using the standard normal probability tables, such confidence intervals can be similarly obtained for any other value of α. These confidence intervals will have confidence coefficient equal to $(1-\alpha)$ when the sampling distribution of the known function $h(\hat{\theta})$ of $\hat{\theta}$ is exactly normal, and it will be approximately $(1-\alpha)$ in other cases.

In case $V(\hat{\theta})$ is not already known, it has to be estimated from the sample data. Let $v(\hat{\theta})$ denote the estimator of $V(\hat{\theta})$. If in the function of $\hat{\theta}$ considered above, $V(\hat{\theta})$ is replaced by $v(\hat{\theta})$, the resulting statistic $(\hat{\theta}-\theta)/\sqrt{v(\hat{\theta})}$ will no longer have standard normal distribution as its sampling distribution. We shall, therefore, have to first determine the sampling distribution and then use it for obtaining confidence interval for θ. Further, the sampling distribution of an estimator differs from estimator to estimator. It also depends on the method of sample selection, sample size, and the distribution of the study variable in the population. It is, therefore, not possible to specify a single sampling distribution for the construction of confidence intervals for all the situations. It is beyond the scope

of this book to work out sampling distribution for every estimator and every sampling situation. However, central limit theorem (Fisz, 1963) ensures that the sampling distribution for most of the estimators $\hat{\theta}$ can be approximated by normal distribution with mean $E(\hat{\theta})$ and variance $V(\hat{\theta})$, when the sample size is large.

The point to be emphasized here is that many estimators we shall use in the text, will not be precisely following normal distribution. However, from Tchebysheff's theorem (Fisz, 1963), at least 75% of the observations from any probability distribution will be within 2 standard deviations of their mean. For the sake of simplicity and to avoid confusion, we shall, therefore, use multiplier 2 in place of 1.96 for building up confidence intervals. Thus, the confidence interval for θ will, in general, be given by

$$\hat{\theta} \pm 2\sqrt{v(\hat{\theta})} \tag{2.8}$$

where $v(\hat{\theta})$ is the estimator of the variance $V(\hat{\theta})$. Such confidence intervals shall provide about .95 confidence coefficient for the estimators $\hat{\theta}$ following approximately normal distribution and a confidence coefficient of at least .75 for any other situation.

The numerical value of the confidence interval in (2.8), for any particular case, will be obtained by using the values of the estimator $\hat{\theta}$ and its variance estimator $v(\hat{\theta})$ computed for that situation.

Example 2.7

In a survey, the sample mean was computed as 796.3, and the value of the variance estimator came out to be 1016.9. Build up the confidence interval for population mean and interpret the results.

Solution

From the statement of the example, $\bar{y}$ =796.3 and $v(\bar{y})$=1016.9. Using relation (2.8), the confidence interval is computed as

$$\bar{y} \pm 2\sqrt{v(\bar{y})}$$

$$= 796.3 \pm 2\sqrt{1016.9}$$

$$= 732.5, 860.1$$

The lower limit 732.5 and the upper limit 860.1, obtained above, provide a reasonable assurance to the investigator that the population mean would lie in the closed interval [732.5, 860.1]. ■

Remark 2.2 To examine the *effect of bias* in the estimator $\hat{\theta}$ on the confidence intervals for the parameter θ, let us assume that the estimator $\hat{\theta}$ is normally distributed with mean $E(\hat{\theta})$ (θ) and variance $V(\hat{\theta})$. Thus, the statistic $Z=[\hat{\theta}-E(\hat{\theta})]/\sqrt{V(\hat{\theta})} = [\hat{\theta}-B(\hat{\theta})-\theta]/\sqrt{V(\hat{\theta})}$, where $B(\hat{\theta})$ is the bias of the estimator $\hat{\theta}$, will be normally distributed with mean zero and variance one. Therefore, the interval

$$[\hat{\theta}-B(\hat{\theta})-Z_{\alpha/2}\sqrt{V(\hat{\theta})},\ \hat{\theta}-B(\hat{\theta})+Z_{\alpha/2}\sqrt{V(\hat{\theta})}] \tag{2.9}$$

will cover parameter θ with probability $(1-\alpha)$. However, the confidence coefficient of the usual confidence interval $[\hat{\theta}-Z_{\alpha/2}\sqrt{V(\hat{\theta})}, \hat{\theta}+Z_{\alpha/2}\sqrt{V(\hat{\theta})}]$, when used in place of the interval in (2.9) for the parameter θ, decreases with the increase in the bias $B(\hat{\theta})$ of the estimator $\hat{\theta}$ (Cochran, 1977).

Remark 2.3 Throughout the book, we shall be using approximately .95 confidence coefficient level for the purpose of illustration. In case it is desired to have confidence intervals for some other confidence level, coefficient 2 in (2.8) is to be replaced by the corresponding standard normal variable value.

Remark 2.4 Some of the situations require to state either a lower or an upper *confidence bound* on a population parameter rather than an interval. For instance, the investigator may wish to state that he/she is reasonably confident that the value of the parameter would not be less than a specified value. In this case, he/she intends to place a lower confidence bound on the parameter. Similarly, if he/she wishes to make a statement that the parameter value is not likely to exceed a specified value, he/she is going to set an upper confidence bound. Lower and upper confidence bounds are determined in the same way as the confidence interval is set. The difference is that the experimenter is providing only one end of the confidence interval. Thus the Z value to be used gets changed, since in this case one does not split α to $\alpha/2$. However, so far as the discussion in this book is concerned, we shall usually confine ourselves to confidence intervals only.

2.7 SAMPLE SIZE DETERMINATION

One of the important aspects in planning a sample survey is to decide about the size of the sample required for estimating the population parameter with a specified precision. The maximum difference between the estimate and the parameter value that can be tolerated on considerations of loss or gain due to policy decisions based on the sample results is termed as *permissible error, tolerable error*, or the *bound on the error of estimation*. Once the permissible error has been specified, the next objective is to determine a sample size that meets these requirements. Since the amount of error differs from sample to sample, the margin of error is specified by the probability statement

$$P[\,|\,\hat{\theta}-\theta\,|\,<B] = 1-\alpha \qquad (2.10)$$

where $(1-\alpha)$ may be taken as 95%, 99%, or some other desired level of confidence, and B is the permissible error. Sometimes, the permissible error is specified in terms of percentage of the value of parameter θ. Such a specification of permissible error can, however, be easily converted into a statement of type (2.10). Theoretically, the required sample size is then determined by equating half width of the confidence interval to the permissible error B, and solving the resulting equation for the sample size n. An analogous approach, for determining the sample size, is followed in case of categorical data. Throughout this book, we shall determine the required sample size by following a two step approach proposed by Stein (1945) and Cox (1952). In the first step, we shall select a small preliminary sample of size n_1. Observations made on the units selected in this sample, will be used to estimate various parameters involved in the expression for the

half width of the confidence interval. After replacing the parameters by their respective estimates obtained from the preliminary sample, half width of the confidence interval will be equated to the permissible error B. The equation is then solved for n, the required sample size. If $n>n_1$, then $(n-n_1)$ additional units are selected, which along with the preliminary sample yield a pooled sample of n units. If $n<n_1$, no more units are selected and preliminary sample is taken as the final sample.

As the expression for estimator of variance (or mean square error) to be used in the confidence interval also varies with the sampling schemes, formulas for determining sample size will change with the sampling procedure. These formulas will, therefore, be presented separately for each sampling and estimation situation.

2.8 SAMPLING AND NONSAMPLING ERRORS

The probability mechanism inherent in the sampling procedure usually selects different units in different samples. The estimates based on the sample observations, as already discussed, will, therefore, differ in general from sample to sample and also from the value of the parameter under consideration. The resultant discrepancy between the sample estimate and the population parameter value is the error of the estimate. Such an error is inherent and unavoidable in any and every sampling scheme, and is termed *sampling error.* This error, however, has the favorable characteristic of being controllable through the size and design of the sample. This kind of error usually decreases with increase in sample size, and shall theoretically become nonexistent in case of complete enumeration. In many situations, the decrease is inversely proportional to the square root of sample size (figure 2.1).

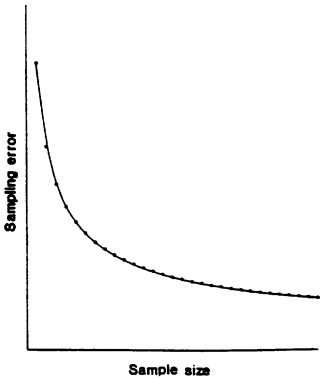

Fig 2.1 Relationship between sampling error and sample size

In any survey study, besides sampling error, there are also errors arising due to defective sampling procedures, ambiguity in definitions, faulty measurement techniques, mistakes in recording, errors in coding-decoding, tabulation and analysis, etc. These errors are known as *nonsampling errors.* For instance, data may be wrongly fed to the computer, or decimal places may be inadvertently changed. If the analysis of data requires transformation of per acre yield to per hectare basis, the multiplier used might be wrong. Sometimes, the respondent may also deliberately respond erroneously while reporting

his answer. A statistician must be aware of this source of error. Unlike the sampling errors, the nonsampling errors are likely to increase with increase in sample size. It is quite possible that nonsampling errors in a complete enumeration survey are greater than both the sampling and nonsampling errors taken together in a sample survey. One should, therefore, be careful in evaluating and checking the processing of the sample data right from its collection to its analysis to minimize the occurrence of nonsampling errors.

LET US DO

2.1 Define expectation and variance of a random variable. Illustrate the definitions with the help of a numerical example.

2.2 From the below given distribution of random variable x, find (a) E(x), (b) E(5x), (c) E(5-x), and (d) $E[x-E(x)]^2$.

x	Probability
11	.08
17	.12
21	.15
24	.30
28	.15
31	.12
49	.08

2.3 A player rolls an unbiased die. If a prime number occurs, he wins an equal number of dollars. Showing up of a nonprime number results in a loss of that number of dollars to him. Is the game favorable to the player ?

2.4 What is a parameter ? Work out mean and variance for the following given population values :
44, 56, 60, 48, 55, 50, 58, 62, 60, 40.

2.5 How does a statistic differ from a parameter ? Explain.

2.6 Explain, in what sense the statistic s^2, the sample mean square, is a random variable?

2.7 What do you understand by the sampling distribution of a statistic ? The knowledge of sampling distribution is important to statistical inference. Explain.

2.8 To what different uses, can the calculated values of sample mean and sample mean square be put ? Discuss.

2.9 Assuming that 20, 12, 15, 16, 18, 14, 22, 28, 24, and 26 are the observations for a sample of 10 units, calculate sample mean and sample mean square.

2.10 Prove algebraically that

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

2.11 What is a point estimator ? Discuss the desirable properties it should possess.

2.12 Five babies were born in a particular year in village Beonhin of Mathura district. The age (in years) of mothers at the time of child birth were 29, 32, 26, 28, and 36. Enumerate all possible WR equal probability samples of size 3, and show numerically that the sample mean age is an unbiased estimator of population mean age of the mothers.

2.13 How does sampling variance differ from mean square error ? State the situation where both are equivalent.

2.14 Distinguish between the population variance and the sampling variance. Illustrate it with a suitable example.

2.15 Describe briefly each of the following terms :

a. Unbiasedness
b. Consistency
c. Relative efficiency
d. Confidence coefficient

2.16 What do you understand by a confidence interval ? Why does an experimenter need it ? Also, discuss the effect of bias in the estimator used to build up the confidence interval for the parameter.

2.17 When we construct a 95% level confidence interval for the population mean, what does it mean ?

2.18 While defining approximately 95% level confidence interval, we have used multiplier 2 in place of exact standard normal value. How will it affect the confidence coefficient for different kinds of sampling distributions ?

2.19 An investigator is to construct confidence interval for population proportion P. Under what circumstances can the normal distribution be used as an approximating sampling distribution for sample proportion p to work out confidence limits ? (*Hint* : For a variable y taking only 0 and 1 values, the sample mean $\bar{y}$ reduces to sample proportion p).

2.20 Why is the determination of sample size important in designing a survey ?

2.21 An investigator has randomly selected 2000 families following WR procedure from a population of 10,000 families. For working out a sufficiently accurate confidence interval for population mean, he/she is to guess the distribution of sample mean in absence of any information regarding the distribution of study variable in the population. Is it reasonable to assume that the sampling distribution is (a) exactly normal, (b) approximately normal, or (c) not at all normal ?

2.22 Distinguish between sampling and nonsampling errors. Which of these errors are more likely to be present in a census or a sample survey ?

CHAPTER 3

Simple Random Sampling

3.1 WHAT IS SIMPLE RANDOM SAMPLING ?

In this book, we shall consider various sampling procedures (schemes) for selection of units in the sample. Since the objective of a survey is to make inferences about the population, a procedure that provides a precise estimator of the parameter of interest is desirable. Many sampling schemes have been developed to achieve this objective. To begin with, simple random sampling, the simplest and the most basic sample selection procedure, is discussed.

Definition 3.1 The sampling procedure is known as *simple random sampling* if every population unit has the same chance of being selected in the sample. The sample thus obtained is termed a *simple random sample.*

For selecting a simple random sample in practice, units from population are drawn one by one. If the unit selected at any particular draw is replaced back in the population before the next unit is drawn, the procedure is called *with replacement* (WR) *sampling*. A set of units selected at n such draws, constitutes a simple random with replacement sample of size n. In such a selection procedure, there is a possibility of one or more population units getting selected more than once. In case, this procedure is continued till n distinct units are selected, and all repetitions are ignored, it is called *simple random sampling* (SRS) *without replacement* (WOR). This method is equivalent to the procedure, where the selected units at each draw are not replaced back in the population before executing the next draw.

Another definition of simple random sampling, both with and without replacement, could be given on the basis of probabilities associated with all possible samples that can be selected from the population.

Definition 3.2 *Simple random sampling* is the method of selecting the units from the population where all possible samples are equally likely to get selected.

3.2 HOW TO DRAW A SIMPLE RANDOM SAMPLE ?

To draw a simple random sample from a population under study is not as trivial as it appears. If the investigator selects the sample by judgement, claiming the sample to be representative of population, it is subjected to investigator bias. Such a sampling leads

to estimators whose properties can not be evaluated. Therefore, one has to use a sampling mechanism which assigns to every population unit an equal probability of being selected in the sample. The most commonly used procedures for selecting a simple random sample are: (1) lottery method, and (2) through the use of random number tables.

3.2.1 *Lottery Method*

In this method, each unit of the population of N units is assigned a distinct identification mark (number) from 1 to N. This constitutes the population frame. Each of these numbers is then written on a different slip of paper. All the N slips of paper are identical in respect of size, color, shape, etc. Fold all these slips in an identical manner and put them in a container or drum, in which a thorough mixing of the slips is carried out before each blindfold draw. The paper slips are then drawn one by one. The units corresponding to the identification labels on the selected slips, are taken to be members of the sample. If the sampling is WR, each slip drawn is put back in the container after noting the identification label on that slip and refolding it. In WOR sampling, the paper slip once drawn is not replaced back in the container. Draws of paper slips, this way, are continued till a sample of required size is obtained.

One can also use a deck of cards, spherical balls or some other such items in place of slips of paper. This procedure of numbering units on slips and selecting slips after reshuffling, becomes tedious when the population size is large. To overcome this difficulty, tables of random numbers are used.

3.2.2 *Through the Use of Random Number Tables*

A *random number table* is an arrangement of ten digits from 0 to 9, occurring with equal frequencies (except for chance fluctuations) independently of each other and without any consistently recurring trends or patterns. Several standard tables of random numbers prepared by Tippett (1927), Fisher and Yates (1938), Kendall and Smith (1939), Rand Corporation (1955), and Rao *et al.* (1974) are available. However, in this book we shall make use of the random number tables due to Rao *et al.* (1974). Some of these random number tables are reproduced in appendix B to help in illustrating the use of random numbers for selecting a sample.

We discuss below three commonly used methods of *using random number tables* for selection of simple random samples.

Direct Approach. Again, the first step in the method is to assign serial numbers 1 to N to the N population units. If the population size N is made up of K digits, then consider K digit random numbers, either row wise or column wise, in the random number table. The sample of required size is then selected by drawing, one by one, random numbers from 1 to N, and including the units bearing these serial numbers in the sample.

This procedure may involve number of rejections of random numbers, since zero and all the numbers greater than N appearing in the table are not considered for selection. The use of random numbers has, therefore, to be modified. Two of the commonly used modified procedures are now discussed.

Remainder Approach. If N is a K digit number, determine the highest K digit multiple of N. Let it be N'. Then a random number r is selected, such that $1 \leq r \leq N'$. The unit bearing the serial number equal to the remainder (say) R, obtained on dividing r by N, is then considered as selected. If remainder is zero, the last unit is selected. As an illustration, let N=24. Here N is a two digit number. The highest two digit multiple of 24 is 96. Let the random number r, selected from 1 to 96, be 83. On dividing 83 by 24, we get remainder as 11. Therefore, the unit bearing serial number 11 is selected in the sample. The process is repeated till the sample of required size is selected. As before, the repeated selections of population units in the sample are permitted for with replacement sample, whereas they are rejected and only distinct units selected for a WOR sample.

Quotient Approach. As before, let N be a K digit number and N' be the highest K digit multiple of N, such that N'= Nm. Select a random number r from 0 to N'-1. Then the unit having serial number (Q+1) is included in the sample, where Q is the quotient when r is divided by m. For instance, if N=24 then N'=96, so that m=4. Let a random number (say) 49 be chosen from 0 to 95. Then Q=12. The unit bearing serial number (Q+1)=13 is then selected in the sample.

It may be noted here, that while using the random number tables, any starting point can be used, and one can move in any predetermined direction along the rows or columns. If more than one sample is to be selected in any problem, each should have its independent starting point.

Besides the above discussed methods, some more methods for sample selection are available in literature. However, being operationally inconvenient, they are usually not employed in practice.

Example 3.1

Appendix C gives data related to the number of tractors in 69 serially numbered villages of Doraha development block in Punjab (India). Select (1) WR and (2) WOR simple random sample of 10 villages using direct approach method.

Solution

Here village is the sampling unit. The villages in the population are already serially numbered which, otherwise, is the first step involved in the sample selection. Refer to appendix B, and use first column by dropping the last two digits of each four digit number. Then we see that the first random number thus formed is 34. Similarly, the subsequent random numbers are seen to be 61,58,...,35.

(1) By selecting the first 10 random numbers from 1 to 69, without discarding repetitions, we obtain the serial numbers of villages in the sample. These are given below along with their variable values (number of tractors).

Village :	34	61	58	62	47	34	11	43	5	35
Tractors :	14	8	15	39	9	14	11	19	12	18

One can see that 34th village has been selected twice in the with replacement simple random sample where repeated selection of units is permitted.

(2) In without replacement sample, any repetition (34th village in the present case) is omitted, and another random number is selected as its replacement. Next random number from 1 to 69 is 26, and it has not appeared earlier. Thus the WOR simple random sample of 10 villages from the population under study is with the following serial numbers :

Village	:	34	61	58	62	47	11	43	5	35	26
Tractors	:	14	8	15	39	9	11	19	12	18	22 ■

3.3 ESTIMATION OF POPULATION MEAN/TOTAL

Once the sample has been selected and the units in the sample observed for the study variable, the next step is to draw inferences about the population from the information contained in the sample. In sample surveys, usually we are interested in estimating certain specific population parameters. The parameters of common interest are mean, total, or the proportion. First we consider equal probability WR sampling.

3.3.1 *WR Simple Random Sampling*

Let y denote the study variable and $Y_1, Y_2,...,Y_N$ be the values of y for N units of the population. Further, let $y_1, y_2,...,y_n$ denote the values of y for n units selected in the sample. Then the sample mean

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

also defined in (2.5), is an unbiased estimator of population mean $\bar{Y}$ (given in 2.3), in case of both with and without replacement simple random sampling. In case of SRS with replacement, the *sampling variance* of mean is given by

$$V(\bar{y}) = \frac{\sigma^2}{n} \tag{3.1}$$

where the population variance σ^2 has already been defined in (2.4).

Variance of the sampling distribution of mean, given above in (3.1), depends on the population parameter σ^2. This value of σ^2 will not be known unless we know all $Y_1, Y_2,...,Y_N$. Since the values of y for all the population units are not known, the actual value of $V(\bar{y})$ can not be obtained. We have, therefore, to satisfy ourselves with only the estimated value of $V(\bar{y})$which we can get from the sample data. An unbiased estimator of $V(\bar{y})$ is then given by

$$v(\bar{y}) = \frac{s^2}{n} \tag{3.2}$$

with the sample mean square s^2 defined in (2.6).

We now summarize the above discussed formulas for the case of WR simple random sampling.

Unbiased estimator of population mean $\bar{Y}$:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad (3.3.)$$

Sampling Variance of $\bar{y}$:

$$V(\bar{y}) = \frac{\sigma^2}{n} \qquad (3.4)$$

Unbiased estimator of $V(\bar{y})$:

$$v(\bar{y}) = \frac{s^2}{n} \qquad (3.5)$$

where σ^2 and s^2 are defined in (2.4) and (2.6) respectively.

In order to clarify the concepts of expected value, variance, and estimator of variance, in relation to simple random sampling WR procedure, we consider an example.

Example 3.2

The height (in cm) of 6 students of M.Sc., majoring in statistics, from Punjab Agricultural University, Ludhiana was recorded during 1985. The data, so obtained, are given below :

Table 3.1 Heights of M.Sc. students

Student	Name	Height
1	Sarjinder Singh	168
2	Gurmeet Singh	175
3	Varinder Kumar	185
4	Sukhjinder Singh	173
5	Devinder Kumar	171
6	Gulshan Kumar	172

1. Calculate (a) population mean $\bar{Y}$, and (b) population variance σ^2.
2. Enumerate all possible SRS with replacement samples of size n=2. Obtain sampling distribution of mean, and hence show that

a. $E(\bar{y}) = \bar{Y}$

b. $V(\bar{y}) = \frac{\sigma^2}{n}$

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

d. $E[v(\bar{y})] = V(\bar{y})$

Solution

(1) Here N=6, and the study variable height is denoted by y. For making computations easily understandable, we present them in tabular form.

Table 3.2 Population data and other computations

Student	Y_i	Y_i^2
1	168	28224
2	175	30625
3	185	34225
4	173	29929
5	171	29241
6	172	29584
Total	1044	181828

It can be easily seen from the above table that

(a) Population mean

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

$$= \frac{1044}{6}$$

$$= 174$$

(b) Population variance

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

$$= \frac{1}{6} [181828 - 6 (174)^2]$$

$$= 28.667$$

(2) Number of all possible samples of size 2, in case of WR sampling, will be $N^n=6^2=36$. The students in various samples along with their corresponding height values, sample means, and sample mean squares are given in table 3.3.

Table 3.3 All possible WR samples and other related statistics

Sample	Sample units	Height of students	$\bar{y}$	s^2
1	(1, 1)	(168, 168)	168.0	0
2	(1, 2)	(168, 175)	171.5	24.5
3	(1, 3)	(168, 185)	176.5	144.5
4	(1, 4)	(168, 173)	170.5	12.5
5	(1, 5)	(168, 171)	169.5	4.5
6	(1, 6)	(168, 172)	170.0	8.0
7	(2, 1)	(175, 168)	171.5	24.5
8	(2, 2)	(175, 175)	175.0	0
9	(2, 3)	(175, 185)	180.0	50.0
10	(2, 4)	(175, 173)	174.0	2.0
11	(2, 5)	(175, 171)	173.0	8.0
12	(2, 6)	(175, 172)	173.5	4.5
13	(3, 1)	(185, 168)	176.5	144.5
14	(3, 2)	(185, 175)	180.0	50.0
15	(3, 3)	(185, 185)	185.0	0
16	(3, 4)	(185, 173)	179.0	72.0
17	(3, 5)	(185, 171)	178.0	98.0
18	(3, 6)	(185, 172)	178.5	84.5
19	(4, 1)	(173, 168)	170.5	12.5
20	(4, 2)	(173, 175)	174.0	2.0
21	(4, 3)	(173, 185)	179.0	72.0
22	(4, 4)	(173, 173)	173.0	0
23	(4, 5)	(173, 171)	172.0	2.0
24	(4, 6)	(173, 172)	172.5	.5
25	(5, 1)	(171, 168)	169.5	4.5
26	(5, 2)	(171, 175)	173.0	8.0
27	(5, 3)	(171, 185)	178.0	98.0
28	(5, 4)	(171, 173)	172.0	2.0
29	(5, 5)	(171, 171)	171.0	0
30	(5, 6)	(171, 172)	171.5	.5
31	(6, 1)	(172, 168)	170.0	8.0
32	(6, 2)	(172, 175)	173.5	4.5
33	(6, 3)	(172, 185)	178.5	84.5
34	(6, 4)	(172, 173)	172.5	.5
35	(6, 5)	(172, 171)	171.5	.5
36	(6, 6)	(172, 172)	172.0	0
Total			6264	1032

Column (4) in table 3.3, lists all possible values of sample mean $\bar{y}$. In case of SRS with replacement, each one of the 36 samples will have equal chance of getting selected and will, therefore, have a probability of $1/N^n=1/36$ associated with it. In case of other WR sampling procedures (other than simple random sampling), these probabilities may not be equal. More appropriately, these can be written in the form of a *sampling distribution* as given in table 3.4. The probability associated with any sample mean value is the number of samples that yield that particular sample mean value, times 1/36.

Table 3.4 Sampling distribution of mean

Serial No.	Sample mean ($\bar{y}$)	Frequency (f)	Probability (p)
1	168.0	1	1/36
2	169.5	2	2/36
3	170.0	2	2/36
4	170.5	2	2/36
5	171.0	1	1/36
6	171.5	4	4/36
7	172.0	3	3/36
8	172.5	2	2/36
9	173.0	3	3/36
10	173.5	2	2/36
11	174.0	2	2/36
12	175.0	1	1/36
13	176.5	2	2/36
14	178.0	2	2/36
15	178.5	2	2/36
16	179.0	2	2/36
17	180.0	2	2/36
18	185.0	1	1/36
Total		36	1

For verifying the other results in the statement, we use table 3.3.

(a) Average of all possible sample means, denoted by $E(\bar{y})$, is obtained from column (4) of table 3.3 as

$$E(\bar{y}) = \frac{1}{36}(168.0+171.5+...+172.0)$$

$$= \frac{6264}{36}$$

$$= 174$$

It is, therefore, seen that

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

Hence $\bar{y}$ is an unbiased estimator of $\bar{Y}$.

(b) By definition 2.7, the variance of all possible sample means in this case, is given by

$$V(\bar{y}) = \frac{1}{36} \sum_{i=1}^{36} \bar{y}_i^2 - [E(\bar{y})]^2$$

On using column (4) of table 3.3, one gets

$$V(\bar{y}) = \frac{1}{36} [(168.0)^2 + (171.5)^2 + ... + (172.0)^2] - (174)^2$$

$$= \frac{1090452}{36} - (174)^2$$

$$= 14.333$$

Now from part (b) of (1), we have

$$\frac{\sigma^2}{n} = \frac{28.667}{2} = 14.333$$

Hence the relation (b) of (2), that is, $V(\bar{y}) = \sigma^2/n$ stands verified.

The reader must note that the $E(\bar{y})$ and $V(\bar{y})$ can also be obtained from the sampling distribution of mean $\bar{y}$, given in table 3.4, by using the expressions

$$E(\bar{y}) = \Sigma\, \bar{y}_i p_i$$

and

$$V(\bar{y}) = \Sigma\, \bar{y}_i^2 p_i - (\Sigma\, \bar{y}_i p_i)^2$$

Here, p_i is the probability of the sample mean $\bar{y}$ taking value $\bar{y}_i$, and the summation is over all the values (18 in this case) taken by $\bar{y}$.
Thus,

$$p_i = \frac{f_i}{N^n}$$

f_i being the number of samples that yield $\bar{y}=\bar{y}_i$.

(c) From table 3.3, we find that the average of all possible sample mean squares is

$$= \frac{1}{36} (0 + 24.5 + ... + 0)$$

$$= \frac{1032}{36}$$

$$= 28.667$$

which equals population variance. It means that

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

This verifies relation (c) of (2).

(d) Just now, we have numerically illustrated that

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

On dividing both sides by n, we get

$$E\left(\frac{s^2}{n}\right) = \frac{\sigma^2}{n}$$

That means, if we work out column s^2/n in table 3.3 and take average of values in that column over all possible samples, it will be same as σ^2/n. Thus $v(\bar{y})$ is seen to be an unbiased estimator of $V(\bar{y})$. These calculations are left as an exercise for the reader. ■

3.3.2 *WOR Simple Random Sampling*

In case of simple random sampling WOR, the sample mean $\bar{y}$ still remains unbiased estimator of the population mean $\bar{Y}$, and has the variance

$$\left.\begin{aligned} V(\bar{y}) &= \frac{N-n}{Nn} S^2 \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \end{aligned}\right] \qquad (3.6)$$

where the *population mean square* S^2 is defined as

$$\left.\begin{aligned} S^2 &= \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \\ &= \frac{1}{N-1} \left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right) \end{aligned}\right] \qquad (3.7)$$

It can be easily seen that the population mean square S^2 and the population variance σ^2 are related through the equation

$$(N-1)\, S^2 = N\sigma^2$$

The variance $V(\bar{y})$ in (3.6) can, therefore, also be written as

$$\left.\begin{aligned} V(\bar{y}) &= \left(\frac{N-n}{N-1}\right) \frac{\sigma^2}{n} \\ &= \left(1 - \frac{n-1}{N-1}\right) \frac{\sigma^2}{n} \end{aligned}\right] \qquad (3.8)$$

All the four forms of $V(\bar{y})$, two in (3.6) and two in (3.8) are equivalent, and one can use any one of them. We shall also not stick to any one particular form of variance $V(\bar{y})$ in the book, and may use any one of the four forms.

As discussed before the actual value of variance in (3.6) or (3.8) can not be found unless we know y values for all the population units. An unbiased estimator of this variance is, therefore, needed and is given by

$$v(\bar{y}) = \frac{N-n}{Nn} s^2$$

where s^2, to be obtained from sample information, is defined earlier in (2.6). Thus for WOR simple random sampling, we have :

Unbiased estimator of $\bar{Y}$ is same as in (3.3).

Variance of estimator $\bar{y}$:

$$V(\bar{y}) = \frac{N-n}{Nn} S^2 \qquad (3.9)$$

Unbiased estimator of $V(\bar{y})$:

$$v(\bar{y}) = \frac{N-n}{Nn} s^2 \qquad (3.10)$$

where S^2 and s^2 are defined in (3.7) and (2.6) respectively.

It can be easily seen that the variance $V(\bar{y})$, in (3.8) or equivalently in (3.9), for WOR case reduces to the variance $V(\bar{y})$ in (3.4) for with replacement sampling when $N=\infty$. Thus, for an infinite population, both with and without replacement sampling procedures are equivalent. Because of this, we call the factor (N-n)/(N-1) as *finite population correction* (fpc).

Remark 3.1 $E(s^2)=\sigma^2$ for SRS with replacement, whereas in case of WOR equal probability sampling we have $E(s^2)=S^2$. For $N=\infty$, fpc takes value 1 whereas the sampling fraction reduces to zero.

Example 3.3

From the data given in example 3.2, enumerate all the SRS without replacement samples of size n=2, and write down sampling distribution of mean. Using this distribution, show that

a. $E(\bar{y}) = \bar{Y}$

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

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

d. $E[v(\bar{y})] = V(\bar{y})$

Solution

Number of possible WOR samples of size n=2 will be $\binom{6}{2}$ =15. The heights of students included in various possible samples, along with sample means ($\bar{y}$) and sample mean squares (s^2), are given in table 3.5.

Table 3.5 All possible WOR samples and other related statistics

Serial No.	Sample units	Height of students	$\bar{y}$	s^2
1	(1 , 2)	(168 , 175)	171.5	24.5
2	(1 , 3)	(168 , 185)	176.5	144.5
3	(1 , 4)	(168 , 173)	170.5	12.5
4	(1 , 5)	(168 , 171)	169.5	4.5
5	(1 , 6)	(168 , 172)	170.0	8.0
6	(2 , 3)	(175 , 185)	180.0	50.0
7	(2 , 4)	(175 , 173)	174.0	2.0
8	(2 , 5)	(175 , 171)	173.0	8.0
9	(2 , 6)	(175, 172)	173.5	4.5
10	(3 , 4)	(185 , 173)	179.0	72.0
11	(3 , 5)	(185 , 171)	178.0	98.0
12	(3 , 6)	(185 , 172)	178.5	84.5
13	(4, 5)	(173 , 171)	172.0	2.0
14	(4 , 6)	(173 , 172)	172.5	.5
15	(5 , 6)	(171, 172)	171.5	.5
Total			2610	516

Column (4) in table 3.5 lists mean values for all possible samples. It can also be written in the form of a *sampling distribution* as shown in table 3.6. Probability (p) values, in this case also, are calculated in the same way as in table 3.4.

Table 3.6 Sampling distribution of mean

Serial No.	Sample mean ($\bar{y}$)	Frequency (f)	Probability (p)
1	169.5	1	1/15
2	170.0	1	1/15
3	170.5	1	1/15
4	171.5	2	2/15
5	172.0	1	1/15
6	172.5	1	1/15
7	173.0	1	1/15
8	173.5	1	1/15

Table 3.6 continued...

Serial No.	Sample mean ($\bar{y}$)	Frequency (f)	Probability (p)
9	174.0	1	1/15
10	176.5	1	1/15
11	178.0	1	1/15
12	178.5	1	1/15
13	179.0	1	1/15
14	180.0	1	1/15
Total		15	1

Using values computed in table 3.5, we proceed to verify the required results.

(a) The average of all possible 15 sample means is given as

$$E(\bar{y}) = \frac{1}{15}(171.5+176.5+...+171.5)$$

$$= \frac{2610}{15}$$

$$= 174$$

which is same as the population mean worked out in example 3.2. Hence $E(\bar{y}) = \bar{Y}$.

(b) By definition 2.7, the variance of $\bar{y}$ is given by

$$V(\bar{y}) = \frac{1}{15}\sum_{i=1}^{15} \bar{y}_i^2 - [E(\bar{y})]^2$$

$$= \frac{1}{15}[(171.5)^2 + (176.5)^2 + ... + (171.5)^2] - (174)^2$$

$$= \frac{454312}{15} - (174)^2$$

$$= 11.467$$

Also, the population mean square

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

$$= \frac{1}{5}[181828 - 6(174)^2]$$

$$= 34.400$$

Then we have

$$\frac{N-n}{Nn}S^2 = \left[\frac{6-2}{(6)(2)}\right](34.400)$$

$$= 11.467$$

This verifies the relation

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

(c) The average of all the 15 sample mean squares from table 3.5 is

$$E(s^2) = \frac{1}{15}(24.5+144.5+...+.5)$$

$$= \frac{516}{15}$$

$$= 34.400$$

which equals population mean square S^2 obtained above in (b). Thus, $E(s^2)=S^2$

(d) In (c) we have seen that

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

On multiplying both sides by (N-n)/Nn, one gets

$$E[v(\bar{y})] = V(\bar{y})$$

which verifies the statement that $v(\bar{y})$ is an unbiased estimator of $V(\bar{y})$. Its numerical verification is left as an exercise for the reader. ■

Example 3.4

From the WOR sample of 10 villages selected in example 3.1, estimate the average number of tractors per village in the block along with its standard error. Also, set up confidence interval for the population mean.

Solution

For convenience, we display the data for 10 selected villages, and the other required computations in table 3.7.

Table 3.7 Sample data and other computations

Village	y	y^2
1	14	196
2	8	64
3	15	225
4	39	1521
5	9	81
6	11	121
7	19	361
8	12	144
9	18	324
10	22	484
Total	167	3521

Our estimate of $\overline{Y}$ average number of tractors per village in the block, is

$$\bar{y} = \frac{1}{10} \sum_{i=1}^{10} y_i$$

$$= \frac{167}{10}$$

$$= 16.7$$

$$\approx 17$$

To find the standard error, we first compute sample mean square as

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

$$= \frac{1}{9} [3521 - 10 (16.7)^2]$$

$$= 81.344$$

The estimate of variance is then given by

$$v(\bar{y}) = \frac{N-n}{Nn} s^2$$

$$= \left[\frac{69-10}{(69)(10)}\right] (81.344)$$

$$= 6.956$$

Estimate of standard error of the sample mean $\bar{y}$ is, therefore,

$$se(\bar{y}) = \sqrt{v(\bar{y})}$$

$$= \sqrt{6.956}$$

$$= 2.637$$

The lower and upper limits of the confidence interval for $\overline{Y}$ are given by

$$\bar{y} \pm 2se(\bar{y})$$
$$= 16.7 \pm 2 (2.637)$$
$$= 11.426, 21.974$$
$$\approx 11, 22$$

Thus, the closed interval [11, 22] covers the average number of tractors per village in the block with probability approximately equal to .95. ■

The estimate of population total Y can be obtained by multiplying the estimate of mean by population size N. Its variance and estimator of variance are N^2 times the corresponding expression for mean $\bar{y}$.

Unbiased estimator of population total Y :

$$\hat{Y} = N\bar{y} \tag{3.11}$$

Variance of estimator $\hat{Y}$:

$$V(\hat{Y}) = N^2V(\bar{y}) \tag{3.12}$$

Unbiased estimator of variance $V(\hat{Y})$

$$v(\hat{Y}) = N^2v(\bar{y}) \tag{3.13}$$

where $V(\bar{y})$ and $v(\bar{y})$ will be used respectively from (3.4) and (3.5) in case of WR sampling, and from (3.9) and (3.10) if the sampling is WOR.

Confidence interval for total Y can also be obtained accordingly following (2.8).

Example 3.5

From the data of WOR sample comprising of 10 villages in example 3.1, estimate the total number of tractors in the development block of 69 villages. Also, set up the confidence interval for it.

Solution

We have N=69 and n=10. For the sake of convenience, the data has been presented in table 3.7 along with some other computations. From (3.11), the estimate of total number of tractors in the block is

$$\begin{aligned}\hat{Y} &= N\bar{y} \\ &= (69)\,(16.7) \qquad \text{(from example 3.4)} \\ &= 1152.3 \\ &\approx 1152\end{aligned}$$

The estimate of variance of $\hat{Y}$ is provided by (3.13). Thus,

$$v(\hat{Y}) = N^2v(\bar{y})$$

Substituting the value of $v(\bar{y})$ from example 3.4, one gets

$$\begin{aligned}v(\hat{Y}) &= (69)^2\,(6.956) \\ &= 33117.5\end{aligned}$$

We now work out confidence interval for the population total. Following (2.8), the required confidence limits will be

$$\hat{Y} \pm 2\sqrt{v(\hat{Y})}$$

$$= 1152.3 \pm 2\sqrt{33117.5}$$

$$= 788.3, 1516.3$$

$$\approx 788, 1516$$

From the limits of the confidence interval obtained above, the investigator is quite confident that the total number of tractors in the block are likely to be in the range of 788 to 1516. ■

The concept of *confidence interval* used in examples 3.4 and 3.5 is further elaborated in the next example.

Example 3.6
Refer to data in appendix C considered for example 3.1. Examine the behavior of approximately 95% level confidence intervals for total number of tractors by selecting 50 WOR simple random samples each of size n=20. Total number of tractors in the block are 1465.

Solution
Fifty different WOR simple random samples of size 20 drawn from the population of 69 villages (appendix C) are presented in appendix D. The estimate $\hat{Y}$ of total and estimated mean square s^2, along with *lower* and *upper confidence limits* (LCL and UCL) computed from these 50 samples using the relation $\hat{Y} \pm 2\sqrt{v(\hat{Y})}$, are exhibited in table 3.8.

Table 3.8 $\hat{Y}$, s^2, LCL, and UCL for n=20 and N=69

Sample	$\hat{Y}$	s^2	LCL	UCL	CI covers Y?
1	1559.4	306.04	1104.5	2014.3	Yes
2	1666.4	381.71	1158.3	2174.4	Yes
3	1607.7	566.64	988.7	2226.7	Yes
4	1373.1	359.67	879.9	1866.3	Yes
5	1197.2	69.29	980.7	1413.6	No
6	1390.4	199.08	1023.4	1757.3	Yes
7	1321.4	129.19	1025.8	1616.9	Yes
8	1183.4	314.34	722.3	1644.4	Yes
9	1566.3	346.33	1082.4	2050.2	Yes
10	1480.1	258.58	1061.9	1898.2	Yes
11	1711.2	565.54	1092.8	2329.6	Yes
12	1666.4	380.77	1158.9	2173.8	Yes
13	1649.1	459.15	1091.9	2206.3	Yes
14	1300.7	187.71	944.4	1656.9	Yes

Table 3.8 continued...

Sample	$\hat{Y}$	s^2	LCL	UCL	CI covers Y?
15	1956.2	536.03	1354.1	2558.2	Yes
16	1680.2	499.92	1098.7	2261.6	Yes
17	1859.6	397.10	1341.4	2377.7	Yes
18	1452.5	329.21	980.6	1924.3	Yes
19	1235.1	307.57	779.1	1691.1	Yes
20	1400.7	328.01	929.7	1871.7	Yes
21	1438.7	248.34	1028.9	1848.4	Yes
22	1276.5	335.21	800.4	1752.6	Yes
23	1518.0	347.16	1033.5	2002.5	Yes
24	1549.1	366.89	1051.0	2047.1	Yes
25	1835.4	362.36	1340.4	2330.4	Yes
26	1280.0	171.63	939.3	1620.6	Yes
27	1814.7	489.48	1239.4	2390.0	Yes
28	1307.6	211.00	929.8	1685.3	Yes
29	1952.7	582.85	1324.9	2580.5	Yes
30	1745.7	406.54	1221.4	2270.0	Yes
31	1335.2	151.71	1014.9	1655.4	Yes
32	1483.5	271.74	1054.8	1912.2	Yes
33	1752.6	571.62	1130.9	2374.3	Yes
34	1518.0	549.89	908.2	2127.8	Yes
35	1500.8	307.57	1044.7	1956.8	Yes
36	1445.6	305.94	990.7	1900.4	Yes
37	1569.8	275.04	1138.5	2001.0	Yes
38	1600.8	346.17	1117.0	2084.6	Yes
39	1169.6	304.89	715.5	1623.6	Yes
40	1649.1	462.62	1089.8	2208.4	Yes
41	1566.3	275.69	1134.5	1998.1	Yes
42	1583.6	234.79	1185.1	1982.0	Yes
43	1656.0	410.63	1129.1	2182.9	Yes
44	1369.7	144.87	1056.7	1682.6	Yes
45	1414.5	368.37	915.4	1913.6	Yes
46	1193.7	132.01	894.9	1492.5	Yes
47	1386.9	76.83	1159.0	1614.8	Yes
48	897.0	39.26	734.1	1060.0	No
49	1411.1	387.31	899.3	1922.8	Yes
50	1145.4	165.94	810.4	1480.4	Yes

This example shows that in about 4% cases the population total Y=1465 falls outside the confidence interval, otherwise, in 96% cases it is covered by the said interval. If many more samples are examined it will tend to reach 95% cases covering the population total. ■

3.4 ESTIMATION OF MEAN/TOTAL USING DISTINCT UNITS

In case of SRS with replacement, the units which get repeated while selecting the sample do not provide any additional information. Therefore, the information obtained from distinct units is sufficient to estimate population mean/total. Let $y_1, y_2, \ldots, y_d$ be the values of the study variable y for the d distinct units in a WR simple random sample of n units. Then, an unbiased estimator of population mean, based on distinct units, is due to Des Raj and Khamis (1958). This estimator, along with its variance and estimator of variance, is given in (3.14), (3.15), and (3.16).

Estimator of population mean $\bar{Y}$ based on distinct units :

$$\bar{y}_d = \frac{1}{d} \sum_{i=1}^{d} y_i \tag{3.14}$$

Variance of estimator $\bar{y}_d$:

$$V(\bar{y}_d) = \left[E\left(\frac{1}{d}\right) - \frac{1}{N} \right] S^2 \tag{3.15}$$

Estimator of variance $V(\bar{y}_d)$:

$$v(\bar{y}_d) = \left[\frac{1}{d} - \frac{1}{N} \right] s_d^2 \tag{3.16}$$

where for $d \geq 2$, $s_d^2 = \frac{1}{d-1} \sum_{i=1}^{d} (y_i - \bar{y}_d)^2$ and S^2 is same as in (3.7).

The estimator $\bar{y}_d$ is always more efficient than the usual WR estimator $\bar{y}$ given in (3.3).

The population total, in this case also, will be estimated by $\hat{Y}_d = N\bar{y}_d$. The expressions for variance $V(\hat{Y}_d)$ and its estimator are, as usual, given by $N^2V(\bar{y}_d)$ and $N^2v(\bar{y}_d)$. For detail, the reader may refer to Murthy (1967).

Example 3.7

Refer to data in part (1) of example 3.1 , where 10 units have been selected in the sample using WR simple random sampling. From observations related to distinct units in this sample, estimate the population total, and also obtain confidence limits for it.

Solution

One can see that in the WR sample of 10 villages drawn in part (1) of example 3.1, the village bearing serial number 34 has been selected twice. On discarding the repetition, the data for 9 distinct villages in the sample is displayed below :

Village	:	34	61	58	62	47	11	43	5	35
Tractors	:	14	8	15	39	9	11	19	12	18

Now we have d = 9 and N = 69. From (3.14), the estimate of the average number of tractors per village in the block is seen to be

$$\bar{y}_d = \frac{1}{d} \sum_{i=1}^{d} y_i$$

$$= \frac{1}{9} (14+8+...+18)$$

$$= 16.11$$

Using the value of $\bar{y}_d$ obtained above, one gets the estimate of the total number of tractors in the block as

$$\hat{Y}_d = N\bar{y}_d$$

$$= 69\ (16.11)$$

$$= 1111.6$$

$$\approx 1112$$

In order to obtain estimated variance of $\hat{Y}_d$, we first compute

$$s_d^2 = \frac{1}{d-1} \sum_{i=1}^{d} (y_i - \bar{y}_d)^2$$

$$= \frac{1}{d-1} \left(\sum_{i=1}^{d} y_i^2 - d\bar{y}_d^2\right)$$

$$= \frac{1}{9-1} [(14)^2 + (8)^2 + ... + (18)^2 - 9(16.11)^2]$$

$$= \frac{1}{8} [3037 - 9(16.11)^2]$$

$$= 87.65$$

From (3.16), the estimated variance of $\hat{Y}_d$ will be

$$v(\hat{Y}_d) = N^2 v(\bar{y}_d)$$

$$= N^2 \left(\frac{1}{d} - \frac{1}{N}\right) s_d^2$$

Making substitutions for different terms, yields

$$v(\hat{Y}_d) = (69)^2 \left(\frac{1}{9} - \frac{1}{69}\right) (87.65)$$

$$= 40319$$

Following (2.8), the confidence limits are obtained as

$$\hat{Y}_d \pm 2\sqrt{v(\hat{Y}_d)}$$

$$= 1111.6 \pm 2\sqrt{40319}$$

$$= 710.0, 1513.2$$

$$\approx 710, 1513$$

The above values indicate that the total number of tractors in 69 villages of Doraha block would probably fall in the closed interval [710, 1513] with probability approximately equal to .95. ■

3.5 DETERMINING SAMPLE SIZE FOR ESTIMATING POPULATION MEAN/TOTAL

So far, in this chapter, we have discussed methods of selecting simple random samples and the point and interval estimation of population mean (or total). The next important topic that merits consideration is the determination of number of units to be included in the sample. If the sample is too large then time, effort, money, and talent are wasted. Conversely, if the number of units included in the sample is too small, we have collected inadequate information which diminishes the utility of the results. This problem can, however, be solved by using the framework of sampling theory.

Though, the required sample size can be determined by using prior information on variance or coefficient of variation of the population (Cochran, 1977; Sukhatme *et al.*, 1984), but usually, it is difficult to obtain reliable information on these parameters. We, therefore, discuss below a two step approach which does not require prior information on the value of any population parameter. Here, a small preliminary sample is used to estimate the population parameter values, which in turn are used to determine final sample size. The preliminary sample is then augmented by drawing additional units from the population, so that, the size of the augmented sample is same as the required final sample size.

Let n_1 be the size of preliminary sample selected using SRS without replacement and

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2$$

where

$$\bar{y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i$$

is the mean of the preliminary sample. Using s_1^2 in place of s^2 in (3.10), and then equating half width of the confidence interval in (2.8) to the permissible error B, one gets the required sample size as

$$n = \frac{Ns_1^2}{ND + s_1^2}$$

where

$$D = \frac{B^2}{4} \tag{3.17}$$

The above equation in case of SRS with replacement reduces to

$$n = \frac{s_1^2}{D}$$

The rule for selecting sample size can then be stated as follows :

Sample size required for estimating population mean with permissible error B :

$$n = \frac{Ns_1^2}{ND + s_1^2} \quad \text{(for SRS without replacement)} \tag{3.18}$$

$$n = \frac{s_1^2}{D} \quad \text{(for SRS with replacement)} \tag{3.19}$$

where D is defined in (3.17). If $n_1 \geq n$, then n_1 is the sufficient sample size and no additional units need to be selected, otherwise, $(n-n_1)$ additional units are to be selected to get the required sample.

Note that, sometimes it is more convenient to express the permissible margin of error as a fraction of true value. In this case $B = \in \overline{Y}$, and in the formulas (3.18) and (3.19), B will have to be replaced by its estimated value $\in \overline{y}_1$.

Example 3.8

An investigator is interested in estimating average number of tractors per village in Doraha development block of Punjab state. The block consists of 69 villages. Using WOR sample of 10 villages selected in example 3.1, as preliminary sample, determine the sample size needed to estimate the said population mean with a margin of error not exceeding 3.

Solution

Here N=69, n_1=10, and B=3. From example 3.4, we have $s_1^2 = 81.344$. Thus, on using (3.17), it is observed that

$$D = \frac{B^2}{4} = 2.25$$

Then from (3.18),

$$n = \frac{Ns_1^2}{ND + s_1^2}$$

$$= \frac{69\,(81.344)}{69\,(2.25) + 81.344}$$

$$= 23.72$$

$$\approx 24$$

It means that the size of preliminary sample is not sufficient for estimating the population mean with desired precision. Therefore, $n-n_1=14$ more villages are required to be selected. ■

In a like manner, we can determine the sample size needed to estimate a population total with desired precision. On equating half width of confidence interval for total, to permissible error and then solving the equation for n, we get the rule for determining sample size analogous to one obtained in (3.18) and (3.19). However, the value of D is different. For reader's convenience, these results are reproduced in (3.20) and (3.21).

Sample size required to estimate population total with permissible error B :

$$n = \frac{Ns_1^2}{ND + s_1^2} \qquad \text{(for SRS without replacement)} \qquad (3.20)$$

$$n = \frac{s_1^2}{D} \qquad \text{(for SRS with replacement)} \qquad (3.21)$$

where $D=B^2/4N^2$. If $n_1 \geq n$, then sample size n_1 is sufficient, otherwise, select $(n-n_1)$ more units to augment the preliminary sample.

Example 3.9

The owner of a poultry farm is interested in estimating the total weight gain, in a period of one month, for N=1500 chicks kept on a new feed. For this purpose, a simple random WOR sample of $n_1=25$ chicks is observed for weight gain. The sample data yielded $s_1^2=45$ gm^2. Determine the sample size required to estimate total weight gain with two kg (=2000 gm) as margin of error.

Solution

In this case, we have N=1500, $n_1=25$, B=2000, $s_1^2=45$, and

$$D = \frac{B^2}{4N^2} = \frac{(2000)^2}{4(1500)^2} = .4444$$

From (3.20), we find that

$$n = \frac{1500\,(45)}{1500\,(.4444) + 45}$$
$$= 94.86$$
$$\approx 95$$

Thus, $n-n_1=95-25=70$ additional chicks are to be selected to get a total sample of required size. ■

3.6 ESTIMATION OF POPULATION PROPORTION

The investigator conducting a sample survey is, sometimes, interested in estimating the *proportion* of the population that possesses a specified attribute. For example, the government might be interested in estimating the proportion of factories using environmental pollution control measures. The interest of the state could also be in estimating the proportion of population in favor of a particular bill, which is to be introduced in parliament for making it a law. A political party might be interested in estimating the proportion of voters likely to vote for it, in the coming elections.

All these examples exhibit a characteristic of dichotomized or binomial population, where an observation either belongs, or does not belong, to the category of interest. The value 1 or 0 is assigned to each unit according as the unit belongs or does not belong to the desired category. The population total for such a variable becomes equal to the number of units in the population possessing the specified attribute. Also, the mean reduces to the proportion P of such units in the population. Similarly, all the other results presented earlier for continuous data, yield corresponding results for this particular case.

Let N_1 units out of N possess the attribute of interest. Then population proportion $P=N_1/N$ and $Q=1-P$. If n_1 units out of n sample units possess this attribute, sample proportion is given by $p=n_1/n$ and $q=1-p$. It can be easily seen that for a random variable y taking values 1 and 0, the sample mean $\bar{y}$ reduces to sample proportion p. Hence, $E(p)=P$ is true for both with and WOR simple random sampling. The important results for the case of estimation of proportion are listed in (3.22) through (3.26).

Unbiased estimator of population proportion P for WR case :

$$p = \frac{n_1}{n} \tag{3.22}$$

Variance of estimator p :

$$V(p) = \frac{PQ}{n} \tag{3.23}$$

Unbiased estimator of variance V(p) :

$$v(p) = \frac{pq}{n-1} \tag{3.24}$$

where $q = 1 - p$, $P = N_1/N$, and $Q = 1 - P$.

Unbiased estimator of P for WOR case is same as in (3.22).

Variance of estimator p :

$$V(p) = \frac{N-n}{N-1}\left(\frac{PQ}{n}\right) \tag{3.25}$$

Unbiased estimator of variance V(p) :

$$v(p) = \left(1-\frac{n}{N}\right)\left(\frac{pq}{n-1}\right) \tag{3.26}$$

For large samples, the sample proportion p can be considered to be approximately normally distributed with mean P and variance V(p). Since the variance V(p) is usually not available in practice, the estimate of variance would be used for setting the confidence interval for P.

Example 3.10

Punjab Agricultural University, Ludhiana, is interested in estimating the proportion P of teachers who consider semester system to be more suitable as compared to the trimester system of education. A with replacement simple random sample of n=120 teachers is taken from a total of N=1200 teachers. The response is denoted by 0 if the teacher does not think the semester system suitable, and 1 if he/she does. From the sample observations given below, estimate the proportion P along with the standard error of your estimate. Also, work out the confidence interval for P.

Teacher :	1	2	3	4	5	6	. . .	119	120	Total
Response :	1	0	1	1	0	1	. . .	0	1	72

Solution

Estimate of P is given by (3.22). Thus,

$$p = \frac{72}{120} = .6$$

Estimate of the standard error of p will then be obtained as

$$\begin{aligned} se(p) &= \sqrt{v(p)} \\ &= \sqrt{\frac{pq}{n-1}} \\ &= \sqrt{\frac{(.6)\,(.4)}{119}} \\ &= .04491 \end{aligned}$$

The confidence limits for P would be obtained following (2.8). Thus,

$$p \pm 2\sqrt{v(p)}$$
$$= .6 \pm 2(.04491)$$
$$= .5102, .6898$$

The proportion of teachers in the university favoring semester system is, therefore, likely to be in the closed interval [.5102, .6898]. ■

In certain situations, the objective could be to estimate the number N_1 of units possessing the attribute of interest. The unbiased estimator $\hat{N}_1$ of N_1 would be N times the sample proportion p defined in (3.22). The variance and its estimator respectively for $\hat{N}_1$, would be N^2 times the variance and the estimator of variance for p.

3.7 SAMPLE SIZE FOR ESTIMATION OF PROPORTION

As in case of quantitative data, let n_1 be the number of units selected in the preliminary sample, and p_1 denote the proportion of units in this sample possessing the attribute under consideration. Also, let $q_1=1-p_1$. Using p_1 and q_1 in place of p and q in (3.26), and then equating half width of confidence interval to the permissible error B, one gets for WOR simple random sampling

$$n = \frac{n_0}{1+(n_0-1)/N}$$

where

$$n_0 = \frac{4p_1q_1}{B^2} + 1 \qquad (3.27)$$

For SRS with replacement case, the above expression for n becomes

$$n = n_0$$

Thus, the formulas for determining the required sample size can be stated as in (3.28) and (3.29).

Sample size required for estimating P with tolerable error B :

In addition to n_1 units included in the preliminary sample, select $n-n_1$ more units, where

$$n = \frac{n_0}{1+(n_0-1)/N} \quad \text{(for SRS without replacement)} \qquad (3.28)$$

$$n = n_0 \quad \text{(for SRS with replacement)} \qquad (3.29)$$

with n_0 defined in (3.27). If $n_1 \geq n$, then n_1 is the required sample size.

Example 3.11
In example 3.10, while estimating P, the investigator feels that the tolerable error could be taken as .08. Do you think the sample size 120 is sufficient ? If not, how many more units should be included in the sample ?

Solution
From example 3.10, we observe that $n_1=120$, $p_1=.6$, and $q_1=.4$. Now from (3.27) and (3.29),

$$\begin{aligned} n &= n_0 \\ &= \frac{4p_1q_1}{B^2} + 1 \\ &= \frac{4(.6)(.4)}{(.08)^2} + 1 \\ &= 151 \end{aligned}$$

The already selected sample of size 120 is thus not sufficient for achieving the given precision. Therefore, $n-n_1=31$ more teachers need to be selected. ■

3.8 ESTIMATION OF PROPORTION USING INVERSE SAMPLING

In practice, a situation may arise, where attribute under consideration prevails with rare frequency. In such cases, the proportion P to be estimated is very small, and estimation procedure described in section 3.6 may not be satisfactory. Even a large sample may not be enough to estimate P with a reasonable degree of precision . The appropriate sample selection procedure for such type of attributes, is known as inverse sampling. The procedure is due to Haldane (1946).

Definition 3.3 The procedure where sampling is continued until a predetermined number of units possessing the attribute are included in the sample, is known as *inverse sampling.*

Let n be the number of units required to be selected to obtain a predetermined number m of units possessing the rare attribute. Though a biased maximum likelihood estimator is also available, we consider an unbiased estimator of P. The variance expressions, for this unbiased estimator, given by Haldane (1946) and Best (1974) are complicated. However, estimator of variance due to Finney (1949) takes a simple form. If the selection of the units is with SRS without replacement, then the unbiased estimator of P in (3.30) follows *negative hypergeometric* distribution. In case of simple random WR selection of units, it is distributed as *negative binomial* (also known as *inverse binomial*).

Unbiased estimator of population proportion P :

$$p = \frac{m-1}{n-1} \tag{3.30}$$

Estimator of variance V(p) :

$$v(p) = \frac{p(1-p)}{n-2}\left(1-\frac{n-1}{N}\right) \quad \text{(when sampling is WOR)} \tag{3.31}$$

$$v(p) = \frac{p(1-p)}{n-2} \quad \text{(when sampling is WR)} \tag{3.32}$$

Example 3.12

A survey conducted by a student of a medical college in Ludhiana town showed that a proportion .008 of adults over 18 years of age, living in a posh colony, are suffering from tuberculosis. Another student of the same college was subsequently given an assignment to examine whether the incidence of tuberculosis infection in the adults of the same age group, living in a slum area, is on the higher side of .008 ? For conducting this survey, voters' lists were used as frame, and voters as the sampling units. It was decided in advance to continue WR simple random sampling of individuals till 10 cases of tuberculosis infection were detected. To arrive at this predetermined number of 10, the investigator had to select 380 adults from the slum area. Besides estimating the proportion in question, work out the confidence limits within which this parameter is expected to lie.

Solution

Here m=10 and n=380. Estimate of proportion of adults suffering from tuberculosis in slum area is, therefore, given by (3.30). Thus,

$$p = \frac{m-1}{n-1}$$

$$= \frac{10-1}{380-1}$$

$$= .02375$$

Proportion of tuberculosis infection among adults of the slum area is approximately three times the proportion of infected adults in the posh colony. We now compute the estimate of variance and the confidence interval using the relations (3.32) and (2.8) respectively. Let us first work out the estimate of variance. From (3.32),

$$v(p) = \frac{p(1-p)}{n-2}$$

$$= \frac{.02375\,(1-.02375)}{378}$$

$$= .000061$$

Confidence limits for the proportion of tuberculosis infected adult population in slum area is obtained as

$$
\begin{aligned}
& p \pm 2\sqrt{v(p)} \\
&= .02375 \pm 2\sqrt{.000061} \\
&= .00809, .03941 \blacksquare
\end{aligned}
$$

3.9 ESTIMATION OVER SUBPOPULATIONS

It may often be impossible to obtain a frame that lists only those units in the population which are of interest. For instance, the investigator is interested in sampling households, where both husband and wife work, or he/she wants to sample households having adults over 50 years of age. However, the best frame available in both the cases is the list of all households in the target area. In this case, before any sample unit is observed, the investigator has no way of knowing whether any particular selected unit is a member of the *subpopulation* under consideration, or not. The procedure for estimating mean or total, therefore, needs to be modified. This modification consists in taking the values of the study variable, for units not belonging to the class of interest, as zero. This indirectly amounts to using only those sample units that belong to the subpopulation of interest.

For further discussion, we shall use the following notations :

- N = the total number of units in the population
- N_1 = the number of units in the subpopulation of interest
- n = the number of units in the WOR simple random sample drawn from the population of size N
- n_1 = the number of units in the sample of size n that belong to the subpopulation under consideration
- Y_{si} = the value of study variable y for the i-th unit of the subpopulation
- y_{si} = the value of y for the i-th sample unit from the subpopulation

The mean, total, and mean square error for the target subpopulation are then given by

$$
\left.
\begin{aligned}
\bar{Y}_s &= \frac{1}{N_1} \sum_{i=1}^{N_1} Y_{si} \\
Y_s &= \sum_{i=1}^{N_1} Y_{si} \\
S_s^2 &= \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} (Y_{si} - \bar{Y}_s)^2
\end{aligned}
\right] \qquad (3.33)
$$

The unbiased estimator of subpopulation mean, and other related results are given in the following box :

Unbiased estimator of the subpopulation mean $\bar{Y}_s$ when N_1 is known :

$$\bar{y}_s = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{si} \tag{3.34}$$

Variance of estimator $\bar{y}_s$:

$$V(\bar{y}_s) = \left[E\left(\frac{1}{n_1}\right) - \frac{1}{N_1}\right] S_s^2 \tag{3.35}$$

Estimator of variance $V(\bar{y}_s)$:

$$v(\bar{y}_s) = \left(\frac{1}{n_1} - \frac{1}{N_1}\right)\left(\frac{1}{n_1 - 1}\right) \left(\sum_{i=1}^{n_1} y_{si}^2 - n_1 \bar{y}_s^2\right) \tag{3.36}$$

In case N_1 is not known, it may be substituted by $n_1 N/n$.

The estimator $\hat{Y}_s$ for the total of the subpopulation is obtained by multiplying the estimator $\bar{y}_s$ by N_1, and the expressions for variance $V(\hat{Y}_s)$ and estimator of variance $v(\hat{Y}_s)$ are arrived at by respectively multiplying $V(\bar{y}_s)$ and $v(\bar{y}_s)$ by N_1^2.

Example 3.13

The family planning wing of the health department of a certain state wishes to conduct a survey at a university campus for estimating the average time gap between the births of children in families having two children. The frame available, of course, lists all the 800 families of the campus. As the prior identification of the families in the population having just two children was difficult, the investigator selected a WOR random sample of 80 families. In the sample families, 32 families were found having two children. These 32 families were interviewed, and the information collected is shown in table 3.9.

Table 3.9 Time gap (in months) between the births of two children

Family	Gap	Family	Gap	Family	Gap	Family	Gap
1	24	9	64	17	57	25	42
2	30	10	32	18	65	26	16
3	50	11	58	19	26	27	37
4	41	12	48	20	35	28	61
5	27	13	51	21	31	29	34
6	47	14	22	22	17	30	29
7	47	15	69	23	28	31	19
8	39	16	54	24	55	32	57

Estimate the average gap between the births of two children, and obtain confidence limits for it.

Solution

We have $n_1=32$, $n=80$, and $N=800$. Let us first work out the estimate of mean given by (3.34). We have

$$\begin{aligned}\bar{y}_s &= \frac{1}{n_1}\sum_{i=1}^{n_1} y_{si} \\ &= \frac{1}{32}(24+30+\ldots+57) \\ &= 41\end{aligned}$$

months as the estimate of average time gap between the births of children. The estimate of variance is provided by (3.36). The expression for it is

$$v(\bar{y}_s) = \left(\frac{1}{n_1} - \frac{1}{N_1}\right)\left(\frac{1}{n_1-1}\right)\left(\sum_{i=1}^{n_1} y_{si}^2 - n_1\bar{y}_s^2\right)$$

Since N_1 is unknown, it would be replaced by

$$\begin{aligned}\hat{N}_1 &= \frac{n_1 N}{n} \\ &= \frac{(32)(800)}{80} \\ &= 320\end{aligned}$$

On substituting $\hat{N}_1$ for N_1, one gets the estimate of variance as

$$\begin{aligned}v(\bar{y}_s) &= \left(\frac{1}{32}-\frac{1}{320}\right)\left(\frac{1}{32-1}\right)[24^2 + 30^2+\ldots+57^2-32(41)^2] \\ &= \frac{(320-32)(233.35)}{(320)(32)} \\ &= 6.56\end{aligned}$$

The required confidence limits are obtained, following (2.8), by using

$$\begin{aligned}&\bar{y}_s \pm 2\sqrt{v(\bar{y}_s)} \\ &= 41 \pm 2\sqrt{6.56} \\ &= 35.88,\ 46.12\end{aligned}$$

The confidence limits obtained above indicate with reasonable confidence that the average gap between the births of two children in the population of families having two children, is likely to be in the range of 35.88 to 46.12 months. ■

The objective in certain situations could be to estimate the proportion of units in a subpopulation possessing a specific attribute. For instance, one may wish to estimate the proportion of men over 70 years of age who are still actively contributing towards family income by way of doing some kind of work (business, farming, job, etc.). However, the frame of such individuals (above 70 years and alive) is not readily available. Instead, a voters' list prepared five years back is available. In this case, the frame

consisting of men expected to cross their 70th year could be prepared from the available voters' list. Since the voters' list was prepared five years back, it could be possible that some of the individuals included in the frame are no more. For a situation like this, the required estimator for the subpopulation proportion and other related expressions can be easily obtained from (3.34), (3.35), and (3.36) by assigning value 1 to the working men who have attained the age of 70, and 0 to the others.

Let us define

$$P_s = \frac{N_1'}{N_1} \text{ and } p_s = \frac{n_1'}{n_1}$$

where n_1' and N_1' are the number of units possessing the attribute of interest out of n_1 and N_1 units respectively.

Unbiased estimator of proportion in the subpopulation when N_1 is known :

$$p_s = \frac{n_1'}{n_1} \tag{3.37}$$

Variance of estimator p_s :

$$V(p_s) = \left[N_1 E\left(\frac{1}{n_1}\right) - 1\right] \frac{P_s(1-P_s)}{N_1 - 1} \tag{3.38}$$

Estimator of variance $V(p_s)$:

$$v(p_s) = \left(1 - \frac{n_1}{N_1}\right) \frac{p_s(1-p_s)}{n_1 - 1} \tag{3.39}$$

Substitute N_1 as $n_1 N/n$ in the above expression, when it is not known.

Example 3.14

Let us consider the example used for discussion above and assume that a sociologist is interested in estimating the proportion of men over 70 years, who are still contributing towards family income by way of doing some work. As mentioned before, the frame is prepared from a five years old voters' list which consists of 1500 men expected to cross their 70th year. Obviously, the frame also includes the names of those voters who expired before, or after reaching the age of 70 years, during the preceding five years period. A sample of n=120 persons was drawn from this frame following WOR simple random sampling. Out of these, 14 individuals were found to have died. On interviewing the remaining 106 persons, it was observed that 21 persons were still actively engaged in earning by doing some kind of work. Estimate the proportion in question, and also obtain the confidence interval for this parameter.

Solution

We are given that N=1500, n=120, n_1=120-14=106, and n_1' = 21. The required estimate of proportion is given by (3.37). Thus,

$$p_s = \frac{n_1'}{n_1} = \frac{21}{106} = .1981$$

We then work out the variance estimator. From (3.39), we have this as

$$v(p_s) = \left(1 - \frac{n_1}{N_1}\right) \frac{p_s (1 - p_s)}{n_1 - 1}$$

Since N_1 is not known, it is estimated by

$$\hat{N}_1 = \frac{n_1 N}{n}$$
$$= \frac{106\,(1500)}{120}$$
$$= 1325$$

On substituting the values of n_1, $\hat{N}_1$ and p_s, in the expression for $v(p_s)$ given above, one gets

$$v(p_s) = \left(1 - \frac{106}{1325}\right)\left[\frac{(.1981)\,(1 - .1981)}{106 - 1}\right]$$
$$= .0014$$

The confidence limits can be worked out following (2.8). These will, therefore, be

$$p_s \pm 2\sqrt{v(p_s)}$$
$$= .1981 \pm 2\sqrt{.0014}$$
$$= .1233,\ .2729$$

The investigator can, therefore, reasonably believe that the proportion of men over 70 years of age, who are actively engaged in supplementing their family income, is likely to be in the closed interval [.1233, .2729]. ■

3.10 SOME FURTHER REMARKS

3.1 A population contains N units, and the value of the study variable y is known for m units of this population. Let these be denoted as $y_1, y_2, \ldots, y_m$. A without replacement simple random sample of n units is selected from the remaining (N-m) units of the population. If $\bar{y}_n$ is the simple mean for the n units selected from the (N-m) units, the estimator

$$\hat{Y}_1 = \sum_{i=1}^{m} y_i + (N - m)\,\bar{y}_n$$

is unbiased for population total Y. Also, it has smaller variance than the estimator $\hat{Y} = N\,\bar{y}$ based on a WOR simple random sample of size n taken from the entire

population. Thus, the advance knowledge of y values for some of the population units can be used profitably.

3.2 In practice, a situation may arise where the estimates obtained from different samples have to be combined to get a pooled estimate of population mean. Let $\bar{y}_1, \bar{y}_2,...,\bar{y}_k$ be the respective means obtained from k independent simple random samples consisting of $n_1, n_2,..., n_k$ units. The minimum variance pooled estimator, based on all the k samples, would be

$$\bar{y}_p = \sum_{i=1}^{k} w_i \bar{y}_i$$

where, for i=1, 2,..., k,

$$w_i = \begin{cases} n_i / \left(\sum_{i=1}^{k} n_i\right) \\ \text{or} \\ \left(\frac{n_i}{N-n_i}\right) / \sum_{i=1}^{k} \left(\frac{n_i}{N-n_i}\right) \end{cases}$$

depending on whether the k samples are drawn using SRS with or without replacement. The variance and estimator of variance for $\bar{y}_p$ are given by

$$V(\bar{y}_p) = \sum_{i=1}^{k} w_i^2 V(\bar{y}_i)$$

$$v(\bar{y}_p) = \sum_{i=1}^{k} w_i^2 v(\bar{y}_i)$$

where the expressions for $V(\bar{y}_i)$ and $v(\bar{y}_i)$ depend on the sampling procedure used for the selection of k samples. The estimator for population total, or proportion, can be obtained in the usual manner from the above estimator of population mean.

LET US DO

3.1 What do you understand by equal probability sampling ? What are the different methods commonly used for selecting a simple random sample ? Keeping in view the population size, which of the methods would you prefer, and why ?

3.2 Is the WOR simple random sample mean unbiased for population mean ? Write the expressions for variance of the sample mean and the unbiased estimator of this variance.

3.3 Using random number tables, select a WR sample of 15 villages from the population of 69 villages in appendix C. From this sample, estimate the total number of tractors in Doraha development block.

3.4 A population consists of 5 M.Sc. students having teaching load of 10, 14, 13,

12, and 15 credit hours in a semester. List all possible WR simple random samples of size 3 that can be drawn from this population. Also, show numerically that

a. $E(\bar{y}) = \bar{Y}$

b. $V(\bar{y}) = \frac{\sigma^2}{n}$

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

3.5 From the population given in exercise 3.4, list all possible WOR samples of size 3, and show numerically that

a. $E(\bar{y}) = \bar{Y}$

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

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

3.6 Refer to data in appendix C, considered for example 3.1. Examine the behavior of the approximately 95% level confidence intervals for total number of tractors by selecting 50 WR simple random samples each of size 25.

3.7 A sociologist wishes to estimate the average age at the time of death for women (age 18 years or more) in a city. The frame used was the list of death records for the year 1992, available in the office of Registrar of Births and Deaths. There were 680 deaths of women during 1992. From these, a WR random sample of 32 deceased women was drawn. The below given information regarding the age of sample women was obtained by contacting their kith and kin. Serial number in the table is the serial number of unit in the population.

Serial No.	Age	Serial No.	Age	Serial No.	Age	Serial No.	Age
102	49	52	66	171	63	259	70
52	66	110	47	54	69	326	44
36	56	211	59	619	46	627	54
447	47	512	60	380	33	280	32
351	71	43	51	210	57	89	50
161	58	14	67	7	67	130	68
7	67	215	61	123	28	431	74
85	74	16	72	54	69	320	60

Estimate the average age at the time of death for women in 1992, and place confidence limits on it.

3.8 During the last decade, some farmers have started raising sunflower crop. Being a new crop, it is grown only by a few farmers in each village. The area under this crop, being quite small, is not entered in the revenue records. The Department of Agriculture is interested in estimating total area under this crop in a district having

900 villages. Since it is difficult to collect information for each village, a WOR simple random sample of 32 villages was drawn. The information collected on area (in hectares) under sunflower cultivation for the sample villages is presented in the table below :

Village	Area	Village	Area	Village	Area	Village	Area
1	2.0	9	0	17	2.0	25	1.0
2	1.5	10	0	18	2.5	26	0
3	1.7	11	1.2	19	1.5	27	1.8
4	2.5	12	1.8	20	1.5	28	0
5	3.5	13	1.0	21	2.0	29	2.4
6	0	14	2.6	22	2.8	30	2.1
7	1.0	15	1.5	23	4.0	31	1.3
8	1.3	16	3.1	24	2.5	32	1.5

Estimate total area under sunflower crop in the district, and place confidence limits on it.

3.9 Refer to data of exercise 3.7, where the sample units have been selected using SRS with replacement method. Making use of observations on distinct units only, estimate the parameter of exercise 3.7, and also obtain confidence interval for it.

3.10 Assume that the sample of 32 women drawn from the population of 680 deceased women in exercise 3.7 is a preliminary sample. Examine, whether this sample size is sufficient to estimate the average age with a margin of error of 5 years? If not, how many more deceased women need to be selected in the sample ?

3.11 A car dealer is feeling concerned over the complaints received in the office of the manufacturer regarding the free service provided by him to the newly purchased cars. To assess the seriousness of the problem, the dealer decided to draw a WR random sample of 70 buyers out of the total of 1400 individuals who had purchased cars through him during the last one year. Twenty one buyers included in the sample graded service provided by him as unsatisfactory. Estimate the percentage of buyers feeling unsatisfied with the service provided, and construct a suitable level confidence interval for it.

3.12 An investigator wishes to estimate the proportion of students in a university whose fathers are graduates. To arrive at the estimate, a WOR simple random sample of 67 students was drawn from a total of 1400 students. On contacting the sampled students, it was found that the fathers of 46 students had not graduated. Estimate the proportion of students whose fathers were at least graduates. Also, set the confidence interval for population proportion.

3.13 Assume the WOR sample of 67 students in exercise 3.12 as the preliminary sample. If the permissible error could be taken as .1, determine how many additional students will have to be selected to estimate the proportion in question with specified precision ?

3.14 The Mayor of a municipal corporation noticed an error in the calculations of general provident fund (GPF) account of an employee. Fearing that such errors might have also crept in the calculations of some other GPF accounts, he directed the audit unit of the corporation to estimate the proportion of such accounts. Expecting that the percentage of such accounts could be quite small, the investigator followed inverse sampling approach. He decided to go on sampling the accounts, using WOR method, from the list of all GPF accounts till 5 wrongly calculated accounts were detected. To arrive at this predetermined number of 5, he had to select 60 accounts out of a total of 1200 GPF accounts. Estimate the proportion of wrongly calculated accounts, and also find the confidence interval for it.

3.15 Earlier, an investigator had estimated the average annual expenses on uniform for school going children, studying in Punjabi (mother tongue) medium government primary schools, in a certain locality. This annual estimated expenses figure was found to be Rs 410. The investigator now wishes to estimate such expenses for children of the same locality studying in English medium public schools. The frame available lists all the families in the locality. The investigator selected 81 families from the population of 700 families. Twenty seven of the sample families were sending their children to English medium public schools. The money spent annually (in rupees) on school uniforms of these children is given below :

Family	Expenses	Family	Expenses	Family	Expenses
1	800	10	1120	19	700
2	1200	11	860	20	960
3	950	12	900	21	850
4	760	13	650	22	600
5	1100	14	750	23	750
6	1050	15	1160	24	800
7	950	16	1000	25	930
8	800	17	900	26	1020
9	1300	18	650	27	730

Estimate the average annual expenses on public school uniforms, and also work out the confidence interval for this average.

3.16 The objective of the survey to be undertaken by the Animal Science department, is to estimate the proportion of families in a town who are selling milk. The frame of families rearing milch cattle is, however, not available. Instead, a list of all the 3350 families in the town is available. To arrive at the estimate, a WOR simple random sample of 360 families was drawn. When contacted, 114 families were found rearing milch cattle. Out of these, 65 families reported that they were selling milk, whereas the others consumed all the milk produced in the family. Estimate the proportion of families selling milk, and place confidence limits on it.

CHAPTER 4

Sampling With Varying Probabilities

4.1 INTRODUCTION

In the preceding chapter, we have discussed simple random sampling in which each unit in the population gets equal chance of being included in the sample. However, when the units vary considerably in size, SRS does not seem to be an appropriate procedure, since it does not take into account the possible importance of the size of the unit. Under such circumstances, selection of units with unequal probabilities may provide more efficient estimators than equal probability sampling. In this scheme, the units are selected with probability proportional to a given measure of size. The size measure is the value of an auxiliary variable (say) x, which is closely associated with the study variable (say) y. This type of sampling is known as *varying probability sampling* or *probability proportional to size* (PPS) *sampling*. For instance, while estimating total number of unemployed youth in a district, the number of households in the village can be used as a size measure when villages are taken as sampling units. Similarly, for estimating total number of tube wells in a certain district, the number of tube wells in a village for a previous period, or net irrigated area for the village, may be taken as size variables.

It may appear that PPS selection procedure would provide biased estimators as the larger units are over represented in the sample. This would be so, if the sample mean is used as an estimator of population mean $\bar{Y}$. However, if the sample observations are appropriately weighted at the estimation stage, taking into consideration their probabilities of selection, it is possible to obtain unbiased estimators. Mahalanobis (1938) was the first to advocate the use of this procedure. Subsequently, it has been discussed in detail by Hansen and Hurwitz (1943), Murthy (1967), Sukhatme *et al.* (1984), and Singh and Chaudhary (1989).

We shall first discuss the methods of selecting a probability proportional to size sample. Thereafter, we shall present some important details of the theory related to PPS sampling with and without replacement for estimating mean/total. The steps involved in each case will be illustrated with suitable examples.

4.2 METHODS OF SELECTING A PPS SAMPLE

The procedure of selecting a PPS sample is based on associating with each unit in the population a set of numbers equal to its size. The selection of units is done corresponding to numbers chosen at random from the totality of numbers so associated with the

population units. The commonly used selection procedures for drawing a PPS sample are (1) cumulative total method, and (2) Lahiri's method. These two selection procedures are now discussed.

4.2.1 *Cumulative Total Method*

Let the size of the i-th unit be denoted by X_i, the total size for N population units being $X = \Sigma X_i$, i=1,2,...,N. Then, the selection procedure consists of following steps :

Steps involved in cumulative total method :

1. Write down cumulative totals for the sizes X_i, i=1,2,...,N.
2. Choose a random number r, such that, $1 \leq r \leq X$.
3. Select i-th population unit if $T_{i-1} < r \leq T_i$, where $T_{i-1} = X_1+X_2+...+X_{i-1}$, and $T_i=T_{i-1} + X_i$.

The probability of selecting the i-th population unit, using this procedure, is given by $P_i = X_i/X$. For selecting a sample of n units with PPS with replacement, the above operation is to be repeated n times.

Example 4.1

A state plans to establish an industrial unit in a particular region. Before doing so, the administration feels it appropriate to estimate total number of unemployed persons in the 14 villages falling within 15 km radius of the proposed site for the factory. The number of households in these 14 villages are 400, 350, 760, 432, 860, 1180 , 530, 600, 1320, 490, 1040, 310, 520, and 900 respectively. Keeping the objective in view, select a sample of 5 villages using varying probability WR sampling, the size being the number of households.

Solution

To make the procedure easily understandable, the required cumulative totals are presented in table 4.1.

Table 4.1 Cumulative totals of sizes $\{X_i\}$

Village	Size (X_i)	Cumulative size (T_i)	Village	Size (X_i)	Cumulative size (T_i)
1	400	400	8	600	5112
2	350	750	9	1320	6432
3	760	1510	10	490	6922
4	432	1942	11	1040	7962
5	860	2802	12	310	8272
6	1180	3982	13	520	8792
7	530	4512	14	900	9692

To select a village, a random number r, $1 \leq r \leq 9692$, is selected by using random number table (see appendix B). Starting with the 11th column consisting of 4 digits, the first random number not exceeding 9692 is 4479. Since $3982 < 4479 \leq 4512$, the 7th unit is, therefore, selected. The next four random numbers to be considered are 191, 8710, 4656, and 8493. Since $0 < 191 \leq 400$, $8272 < 8710 \leq 8792$, $4512 < 4656 \leq 5112$, and $8272 < 8493 \leq 8792$, the 1st, 13th, 8th, and 13th villages are selected. The 13th village is selected twice. Hence, the required sample, selected using PPS with replacement, will contain the villages bearing serial numbers 1, 7, 13, 8, and 13. ■

4.2.2 *Lahiri's Method*

We have noticed that the cumulative total method involves cumulation of sizes and then writing down these cumulative totals. This step becomes quite tedious when N is large. A procedure which avoids the need for calculating cumulative totals for each unit has been given by Lahiri (1951). It involves the following steps for selecting a sample:

Steps in sample selection through Lahiri's method :

1. Select a random number (say) i from 1 to N.
2. Select another random number (say) j, such that $1 \leq j \leq M$, where M is either equal to the maximum of the sizes $\{X_i\}$, i= 1,2, ..., N, or is more than the maximum size in the population.
3. If $j \leq X_i$, the i-th unit is selected, otherwise, the pair (i, j) of random numbers is rejected, and another pair is chosen by repeating the steps (1) and (2).

For selecting a sample of n units, the procedure is to be repeated till n units are selected. In cases where the maximum size in the population can not be determined, M is taken so large that it will surely be more than the maximum size. However, a very large value of M may result in an increase in the number of rejected pairs (i , j).

Example 4.2

Select a WR varying probability sample of size 2 from the population in example 4.1, by using Lahiri's method.

Solution

Here N=14, n = 2, and M = 1320. If this value of M was not available, one could take M equal to 2000, or 3000, or even 10,000. First, we have to select a random number i, $1 \leq i \leq 14$, and then a second random number j, $1 \leq j \leq 1320$, is to be selected. Referring to the random number table and using the first two digits of column 11 for selecting i and all the 4 digits of column 12 for selecting j, we get the selected pair as (1, 1282). Since $1282 > X_1$ (=400), the first pair of random numbers is rejected. Similarly, we find that the next pair of random numbers (12, 853) is also rejected while the third pair (11, 47) results in the selection of 11th village in the sample since $47 < X_{11}$(=1040). Proceeding further, we find that the next three pairs of random numbers (6, 1265), (11, 1056), and (1, 812) also get rejected, while the fourth pair (8, 490) results in the selection of 8th village since $490 < X_8$ (=600). The required sample selected by Lahiri's method will, therefore, include the villages bearing serial numbers 8 and 11. ■

4.3 ESTIMATION IN PPSWR SAMPLING

Consider a population U= $(U_1, U_2,...,U_N)$ consisting of N distinct and identifiable units. Associated with each U_i, i=1, 2,...,N, are two values - Y_i, of study variable, and X_i, of auxiliary variable. It is assumed that X_i's are known for all i and $X = \Sigma X_i$, so that, $P_i = X_i/X$, i=1,2,...,N. Let (y_i, p_i), i=1,2,...,n, be the values of the study variable and respective probabilities of selection for units included in the sample. Then important results pertaining to estimation of population total $Y = \Sigma Y_i$, i=1,2,...,N, are given below:

Unbiased estimator of population total Y:

$$\hat{Y}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i} \tag{4.1}$$

Variance of estimator $\hat{Y}_{pps}$:

$$\left.\begin{aligned} V(\hat{Y}_{pps}) &= \frac{1}{n} \left(\sum_{i=1}^{N} \frac{Y_i^2}{P_i} - Y^2\right) \\ &= \frac{1}{n} \sum_{i=1}^{N} \left(\frac{Y_i}{P_i} - Y\right)^2 P_i \end{aligned}\right] \tag{4.2}$$

with $P_i = X_i/X$, i=1, 2,..., N.

Unbiased estimator of variance $V(\hat{Y}_{pps})$:

$$\left.\begin{aligned} v(\hat{Y}_{pps}) &= \frac{1}{n(n-1)} \left(\sum_{i=1}^{n} \frac{y_i^2}{p_i^2} - n\hat{Y}^2_{pps}\right) \\ &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{Y}_{pps}\right)^2 \end{aligned}\right] \tag{4.3}$$

The estimator for population mean $\overline{Y}$ can simply be obtained by dividing $\hat{Y}_{pps}$ by N. The corresponding variance and estimator of variance can be easily obtained by dividing $V(\hat{Y}_{pps})$ and $v(\hat{Y}_{pps})$ respectively by N^2. Thus we have the following results :

Unbiased estimator of population mean $\overline{Y}$:

$$\bar{y}_{pps} = \frac{1}{nN} \sum_{i=1}^{n} \frac{y_i}{p_i} \tag{4.4}$$

Variance of estimator $\bar{y}_{pps}$:

$$\left.\begin{aligned} V(\bar{y}_{pps}) &= \frac{1}{nN^2} \left(\sum_{i=1}^{N} \frac{Y_i^2}{P_i} - Y^2\right) \\ &= \frac{1}{nN^2} \sum_{i=1}^{N} \left(\frac{Y_i}{P_i} - Y\right)^2 P_i \end{aligned}\right] \tag{4.5}$$

Unbiased estimator of variance $V(\bar{y}_{pps})$:

$$v(\bar{y}_{pps}) = \frac{1}{n(n-1)N^2}\left(\sum_{i=1}^{n}\frac{y_i^2}{p_i^2} - n\hat{Y}_{pps}^2\right)$$

$$= \frac{1}{n(n-1)N^2}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_{pps}\right)^2 \qquad (4.6)$$

Example 4.3

The data relating to irrigated area of 69 villages of Doraha development block are given in appendix C. Select a sample of 10 villages, using PPS with replacement sampling, taking net irrigated area as the size measure. Therefrom, estimate the total number of tube wells in the block along with its standard error, and place confidence limits on the population total.

Solution

From appendix C, we find that X =23857. A with replacement PPS sample of n=10 villages can be selected by using either the cumulative total method, or the method due to Lahiri (1951). Both these methods are discussed in section 4.2. Let us assume that by using one of these methods, we get the sample of villages with serial numbers 8, 38, 35, 19, 36, 8, 28, 68, 9, and 53. Study variable values in respect of the selected villages are given below :

Village	Tube wells	Village	Tube wells
8	70	8	70
38	97	28	118
35	116	68	131
19	551	9	219
36	115	53	100

For easy understanding of various steps involved in the estimation of total number of tube wells, we display the required computations in table 4.2.

Table 4.2 Sample data and other required computations

Village	Tube wells (y_i)	Net irrigated area (x_i)	y_i/x_i	$y_i^2/x_i^2 = (y_i/x_i)^2$	y_i^2/x_i
8	70	180	.3889	.1512	27.22
38	97	467	.2077	.0431	20.15
35	116	346	.3353	.1124	38.89
19	551	1178	.4677	.2187	257.73
36	115	261	.4406	.1941	50.67
8	70	180	.3889	.1512	27.22
28	118	429	.2751	.0757	32.46
68	131	269	.4870	.2372	63.80
9	219	458	.4782	.2287	104.72
53	100	284	.3521	.1240	35.21
Total			3.8215	1.5363	658.07

On using (4.1), the estimate of total number of tube wells in Doraha block can be obtained as

$$\hat{Y}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}$$

$$= \frac{X}{n} \sum_{i=1}^{n} \frac{y_i}{x_i}$$

$$= \frac{23857\,(3.8215)}{10}$$

$$= 9116.95$$

Using (4.3), the estimator of variance $V(\hat{Y}_{pps})$ thus becomes

$$v(\hat{Y}_{pps}) = \frac{1}{n\,(n-1)} \left(\sum_{i=1}^{n} \frac{y_i^2}{p_i^2} - n\,\hat{Y}_{pps}^2 \right)$$

$$= \frac{1}{n\,(n-1)} \left(X^2 \sum_{i=1}^{n} \frac{y_i^2}{x_i^2} - n\hat{Y}_{pps}^2 \right)$$

$$= \frac{1}{10\,(9)} \left[(23857)^2\,(1.5363) - 10\,(9116.95)^2 \right]$$

$$= 480080.88$$

Then,

$$se\,(\hat{Y}_{pps}) = \sqrt{480080.88}$$

$$= 692.88$$

Following (2.8), the confidence limits for population total are obtained as

$$\hat{Y}_{pps} \pm 2\,se(\hat{Y}_{pps})$$

$$9116.95 \pm 2(692.88)$$

$$= 7731.19,\ 10502.71$$

After rounding off to whole numbers, the end points of the confidence interval become (7731, 10503). ■

4.4 RELATIVE EFFICIENCY OF PPSWR ESTIMATOR

The PPS sampling is expected to be more efficient than SRS when the size measure x is approximately proportional to the study variable y, and the line of regression of y on x passes through, or nearly through, the origin. The percent relative efficiency of the PPS estimator $\hat{Y}_{pps}$ in relation to the usual SRS with replacement based estimator $\hat{Y}$, in (3.11), is given by

$$RE = \frac{V(\hat{Y})}{V(\hat{Y}_{pps})}(100) \tag{4.7}$$

where the variances $V(\hat{Y})$ and $V(\hat{Y}_{pps})$ are respectively given in (3.12) and (4.2). The steps involved in computation of the said relative efficiency are illustrated by taking a suitable population data.

Example 4.4

The data related to number of agricultural laborers (y) in 1992 and total population (x) in 1981 for 24 villages of Samrala development block are given in the following table. The letters y and x stand for study and auxiliary variables.

Table 4.3 Data on agricultural laborers and total population

y	x	y	x	y	x
713	2442	252	923	78	215
98	368	147	526	217	773
68	217	346	1121	423	1495
143	498	835	2797	230	803
311	1108	1002	3264	779	2586
173	706	384	1244	361	1141
356	1190	574	1911	205	1042
295	1046	95	427	193	721

Work out the relative efficiency of WR varying probabilities sampling based estimator $\hat{Y}_{pps}$ in relation to the usual SRS with replacement estimator $\hat{Y}$ given in (3.11), for the total number of agricultural laborers if a sample of size 6 is to be drawn in both the cases.

Solution

In this case, we have N=24 and n=6. First of all, we obtain population totals Y and X. Thus,

$$\begin{aligned} Y &= 713 + 98 + ... + 193 \\ &= 8278 \\ X &= 2442 + 368 + ... + 721 \\ &= 28564 \end{aligned}$$

Then, using (2.4), we compute population variance for variable y as

$$\begin{aligned} \sigma^2 &= \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \overline{Y}^2 \\ &= \frac{1}{24}[(713)^2 + (98)^2 + ... + (193)^2] - \left(\frac{8278}{24}\right)^2 \\ &= 63403.7 \end{aligned}$$

Also,

$$\sum_{i=1}^{N} \frac{Y_i^2}{P_i} - Y^2 = X \sum_{i=1}^{N} \frac{Y_i^2}{X_i} - Y^2$$

$$= (28564)\left[\frac{(713)^2}{2442} + \frac{(98)^2}{368} + \ldots + \frac{(193)^2}{721}\right] - (8278)^2$$

$$= (28564)\,(2417.3153) - (8278)^2$$

$$= 522910.2$$

Now from (3.12), the variance $V(\hat{Y})$ for SRS with replacement case is given by

$$V(\hat{Y}) = \frac{N^2\sigma^2}{n}$$

On making substitutions, one gets

$$V(\hat{Y}) = \frac{(24)^2\,(63403.7)}{6}$$

$$= 6086755.2$$

In case of PPS with replacement, the sampling variance of estimator $\hat{Y}_{pps}$ is given by (4.2). Therefore,

$$V(\hat{Y}_{pps}) = \frac{1}{n}\left(\sum_{i=1}^{N} \frac{Y_i^2}{P_i} - Y^2\right)$$

$$= \frac{522910.2}{6}$$

$$= 87151.7$$

The percent relative efficiency of estimator $\hat{Y}_{pps}$ in relation to estimator $\hat{Y}$ will, therefore, be

$$RE = \frac{V(\hat{Y})}{V(\hat{Y}_{pps})}(100)$$

$$= \frac{6086755.2}{87151.7}(100)$$

$$= 6984.1$$

It shows that the PPS based estimator $\hat{Y}_{pps}$, for the given population, is 69.841 times more efficient than the one based on SRS. ■

Notice that, for determining the relative efficiency of the estimator $\hat{Y}_{pps}$ with respect to the SRS with replacement estimator $\hat{Y}$, information on the study and auxiliary variables for all the population units has been used. In practice, the information on the study

variable y will not be available for all the population units. It will be available only for the units in the sample. Therefore, if the investigator wishes to determine whether the use of auxiliary information for the selection of sample has been beneficial, or not, the relative efficiency has to be estimated from sample data. For this purpose, one will need estimates of the variances $V(\hat{Y}_{pps})$ and $V(\hat{Y})$ from the selected sample.

The unbiased estimator of $V(\hat{Y}_{pps})$ can be had from (4.3), whereas an unbiased estimator of variance for the SRS based estimator $\hat{Y}$, from a PPS with replacement sample, is given by (4.8) below :

$$v_{pps}(\hat{Y}) = \frac{1}{n^2}\left(N \sum_{i=1}^{n} \frac{y_i^2}{p_i} - n\,\hat{Y}_{pps}^2\right) + \frac{v(\hat{Y}_{pps})}{n} \qquad (4.8)$$

On using (4.3) and (4.8), we obtain expression for the estimator of relative efficiency of varying probability WR sampling in relation to SRS with replacement. Thus we get,

$$RE = \frac{v_{pps}(\hat{Y})}{v(\hat{Y}_{pps})}(100) \qquad (4.9)$$

How to evaluate the above estimated relative efficiency from sample data, is illustrated through the following example.

Example 4.5

From the sample data obtained for the problem considered in example 4.3, estimate the relative efficiency of varying probability WR sampling in relation to SRS with replacement.

Solution

From example 4.3, we have N=69, n=10, X=23857, $\hat{Y}_{pps}$=9116.95, and $v(\hat{Y}_{pps})$=480080.88. In order to estimate percent relative efficiency of PPSWR sampling over SRS with replacement, from the selected PPS sample, we are to obtain the value of $v_{pps}(\hat{Y})$. For this, let us compute

$$\sum_{i=1}^{n} \frac{y_i^2}{p_i} = X \sum_{i=1}^{n} \frac{y_i^2}{x_i}$$

On using the value of $\Sigma\, y_i^2/x_i$, i=1,2,...,n, from the last column of table 4.2, and population total X=23857, one gets

$$\sum_{i=1}^{n} \frac{y_i^2}{p_i} = 23857\ (658.07)$$
$$= 15699575$$

Now the expression for $v_{pps}(\hat{Y})$ from (4.8) is

$$v_{pps}(\hat{Y}) = \frac{1}{n^2}\left(N \sum_{i=1}^{n} \frac{y_i^2}{p_i} - n\,\hat{Y}_{pps}^2\right) + \frac{v(\hat{Y}_{pps})}{n}$$

On making substitutions of different values already obtained, it yields

$$v_{pps}(\hat{Y}) = \frac{1}{(10)^2}[(69)(15699575) - (10)(9116.95)^2] + \frac{480080.88}{10}$$

$$= 2568837.1$$

Then from (4.9), we get the required estimate of percent relative efficiency as

$$RE = \frac{v_{pps}(\hat{Y})}{v(\hat{Y}_{pps})}(100)$$

$$= \frac{2568837.1}{480080.88}(100)$$

$$= 535.08 \blacksquare$$

4.5 DETERMINING SAMPLE SIZE FOR ESTIMATING POPULATION MEAN/TOTAL

Let n_1 be the number of units selected in the preliminary sample, and

$$s_{z1}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(\frac{y_i}{p_i} - \hat{Y}_{pps1}\right)^2 \qquad (4.10)$$

where

$$\hat{Y}_{pps1} = \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{y_i}{p_i}$$

is the preliminary sample estimate of population total. Use s_{z1}^2/n in place of $v(\hat{Y}_{pps})$, and then equate half width of the confidence interval to the permissible error B. Solution of this equation for n, yields (4.11). Analogously, one can arrive at (4.12) which gives the sample size needed to estimate a population mean. The rule for selecting sample size then becomes as follows :

Sample size required for estimation of population total/mean with permissible error B :

$$n = \frac{4s_{z1}^2}{B^2} \qquad \text{(for population total)} \qquad (4.11)$$

$$n = \frac{4s_{z1}^2}{N^2B^2} \qquad \text{(for population mean)} \tag{4.12}$$

where s_{z1}^2 has been defined in (4.10). If $n_1 \geq n$, then n_1 is the required sample size, otherwise, $(n-n_1)$ more units are to be selected to augment the initial sample of n_1 units.

Example 4.6

A state government is interested in estimating the total number of tube wells in Doraha development block of Punjab. The block consists of 69 villages. Using PPS with replacement sample selected in example 4.3 as the preliminary sample, determine the sample size required to estimate the said population total with a margin of error not exceeding 900 tube wells. Take approximate confidence coefficient as .95.

Solution

Here N=69, n_1=10, and B=900. From example 4.3, we have s_{z1}^2 = (10) v $(\hat{Y}_{pps})$ = 10(480080.88). Thus, from (4.11) we find

$$n = \frac{4s_{z1}^2}{B^2}$$

$$= \frac{4\,(10)\,(480080.88)}{(900)^2}$$

$$= 23.7$$

which on rounding off to whole number, becomes 24. Since 24>n_1 (= 10), the investigator will have to select (24-10)=14 more units for estimating total number of tube wells, with 900 tube wells as the margin of error. ■

4.6 SAMPLING WITH PPS WITHOUT REPLACEMENT

As in case of simple random sampling, the sample units can be selected without replacement in PPS sampling also. The estimators of population mean/total in PPS without replacement case are generally more efficient than those in with replacement sampling, as the effective sample size is more in the former case. However, the sampling procedure and the corresponding theory becomes more complicated.

Assuming one by one selection of units for the easiest case of a sample of size n=2, the probabilities of selecting i-th, j-th, and the pair of i-th and j-th units in the sample are given by

$$\pi_i = P_i \left(\sum_{j=1}^{N} \frac{P_j}{1-P_j} + 1 - \frac{P_i}{1-P_i} \right) \tag{4.13}$$

$$\pi_j = P_j \left(\sum_{i=1}^{N} \frac{P_i}{1-P_i} + 1 - \frac{P_j}{1-P_j} \right) \tag{4.14}$$

$$\pi_{ij} = P_i P_j \left(\frac{1}{1-P_i} + \frac{1}{1-P_j} \right) \tag{4.15}$$

where $\{P_i\}$ are the initial selection probabilities. Since the probability of selecting a particular population unit changes from draw to draw, it makes the procedure complicated for $n > 2$. A number of procedures have been proposed to select samples of size greater than two, by various workers. However, most of these selection procedures and corresponding theory are complicated, and not easily applicable in practice. Keeping the scope of the book in view, we shall confine ourselves to some well known estimators and sampling methods.

Sections 4.7, 4.8, and 4.9 of this chapter discuss three important estimators due to Des Raj (1956), Murthy (1957), and Horvitz and Thompson (1952), whereas in sections 4.10 and 4.11 are discussed two practically convenient selection procedures. The estimator, considered in section 4.7, is based on order of selection of units in the sample. The corresponding unordered estimator, which does not depend on the order of selection of units, is discussed in section 4.8. The estimator in section 4.9 is general, and can be used in all PPS without replacement sampling cases, where inclusion probabilities for different units in the sample can be calculated.

4.7 DES RAJ'S ORDERED ESTIMATOR

The estimators that take into account the order in which the units are selected in the sample are called *ordered estimators.* Some such estimators have been proposed by Das (1951), Sukhatme (1953), and Des Raj (1956). Des Raj considers a sample of size n drawn WOR with probabilities of selection at each draw being proportional to $P_i = X_i/X$, $i = 1, 2, ..., N$. Let $u_1, u_2, ..., u_n$, be the units selected in order of their draw, and let $\{y_i\}$ and $\{p_i\}$ be the corresponding sets of y-values and their initial probabilities of selection. Then we have :

Des Raj's ordered unbiased estimator of population total Y for a sample of size n :

$$\hat{Y}_d = \frac{1}{n} \sum_{i=1}^{n} t_i \tag{4.16}$$

where

$$t_1 = \frac{y_1}{p_1}$$

$$t_i = y_1 + y_2 + ... + y_{i-1} + \frac{y_i}{p_i}(1 - p_1 - p_2 - ... - p_{i-1}), \quad (i = 2,3,...,n)$$

Unbiased estimator of variance $V(\hat{Y}_d)$:

$$v(\hat{Y}_d) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left[t_i - \frac{1}{n} \left(\sum_{i=1}^{n} t_i \right) \right]^2$$

$$= \frac{1}{n(n-1)} \left(\sum_{i=1}^{n} t_i^2 - n\hat{Y}_d^2 \right) \tag{4.17}$$

The expression for variance $V(\hat{Y}_d)$ being complicated, has not been included in the list of formulas above. The estimator $\hat{Y}_d$ is more efficient than the usual WR estimator $\hat{Y}_{pps}$, and the variance estimator $v(\hat{Y}_d)$ is always positive.

For n=2, the estimator $\hat{Y}_d$ and variance estimator $v(\hat{Y}_d)$ take simpler forms.

Des Raj's ordered unbiased estimator of population total Y for a sample of size two :

$$\hat{Y}_d = \frac{1}{2}\left[\frac{y_1(1+p_1)}{p_1} + \frac{y_2(1-p_1)}{p_2}\right] \tag{4.18}$$

Estimator of variance $V(\hat{Y}_d)$:

$$v(\hat{Y}_d) = \frac{1}{4}(1-p_1)^2\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2 \tag{4.19}$$

Example 4.7

The people of a particular area in Samrala development block of Ludhiana district requested the government to set up a veterinary hospital in their area. In order to plan the hospital's requirements in respect of staff and facilities, the state thinks it appropriate to have an idea about the total number of pet animals in this area consisting of 14 villages. For this purpose, select a PPS without replacement sample of size 2 using Lahiri's method taking number of households (table 4.4) in the village as the size variable. Therefrom, estimate total number of pet animals in the area along with its standard error by using Des Raj's ordered estimator. Also, place the confidence limits on the population total.

Solution

For sake of convenience, the values of $\{P_i\}$ and $\{1-P_i\}$ that will be required in subsequent discussion are also given in table 4.4.

Table 4.4 Values of Y_i , X_i, P_i, and $(1-P_i)$

Village No.	Village name	Pet animals (Y_i)	Households (X_i)	$P_i=X_i/X$	$1-P_i$
1	Jatana	740	274	.0726	.9274
2	Jaspalon	690	248	.0657	.9343
3	Chak Sarai	520	194	.0514	.9486
4	Khehra	490	161	.0427	.9573
5	Tamkodi	185	62	.0164	.9836
6	Rampur	1880	988	.2618	.7382
7	Balala	345	178	.0472	.9528
8	Mehdoodan	240	111	.0294	.9706
9	Madpur	680	309	.0819	.9181
10	Rupalon	450	210	.0556	.9444

Table 4.4 continued...

Village No.	Village name	Pet animals (Y_i)	Households (X_i)	$P_i=X_i/X$	$1-P_i$
11	Lopon	435	209	.0554	.9446
12	Lalkalan	470	221	.0586	.9414
13	Dhindsa	375	162	.0429	.9571
14	Begowal	865	447	.1184	.8816
Total		Y=8365	X=3774		

For selecting units in the sample, we choose a pair of random numbers (i, j) such that $1 \le i \le 14$ and $1 \le j \le 988$. We get this pair as (9, 379). Since $379 > X_9$ (=309), this pair of random numbers is rejected. Again we select another pair of random numbers (i, j) as (2, 78). Now $78 \le X_2$ (=248), thus the village bearing serial number 2 is selected in the sample.

Since the sampling procedure is WOR, the selected unit bearing serial number 2 is not replaced back. The number of units left in the population now are 13. We are to select another unit from these remaining 13 villages. For this purpose, we modify table 4.4 by deleting selected second village from the list of 14 villages, and giving new serial numbers to the remaining 13 villages. This yields us the table 4.5.

Table 4.5 Values of Y_i and X_i after deleting the village number 2

Village No.	Village name	Pet animals (Y_i)	Households (X_i)
1.	Jatana	740	274
2.	Chak Sarai	520	194
3.	Khehra	490	161
4.	Tamkodi	185	62
5.	Rampur	1880	988
6.	Balala	345	178
7.	Mehdoodan	240	111
8.	Madpur	680	309
9.	Rupalon	450	210
10.	Lopon	435	209
11.	Lalkalan	470	221
12.	Dhindsa	375	162
13.	Begowal	865	447

Now, for selecting the next unit in the sample, we choose another pair of random numbers (i, j) such that $1 \le i \le 13$ and $1 \le j \le 988$. Let the chosen pair be (8, 129). As $129 \le X_8$ (=309), the 8th village is selected in the sample. The serial number of this selected village, in the original table 4.4, is 9. Thus, the sample selected consists of villages bearing serial numbers 2 and 9. On observation during the survey, these selected villages will be found to have the number of pet animals as 690 and 680. Now, Des Raj's ordered estimator of Y is

$$\hat{Y}_d = \frac{1}{2}\left[\frac{y_1}{p_1}(1+p_1) + \frac{y_2}{p_2}(1-p_1)\right]$$

$$= \frac{1}{2}\left[\frac{690}{.0657}(1+.0657) + \frac{680}{.0819}(1-.0657)\right]$$

$$= 9474.8$$

$$\approx 9475$$

The estimator of variance from (4.19) is

$$v(\hat{Y}_d) = \frac{1}{4}(1-p_1)^2\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2$$

On substituting different values, it yields

$$v(\hat{Y}_d) = \frac{1}{4}(1-.0657)^2\left(\frac{690}{.0657} - \frac{680}{.0819}\right)^2$$

$$= 1055724.7$$

The estimated standard error of $\hat{Y}_d$ will, therefore, be

$$se(\hat{Y}_d) = \sqrt{1055724.7}$$

$$= 1027.5$$

The confidence limits, for the total number of pet animals in 14 villages, are obtained as

$$\hat{Y}_d \pm 2\sqrt{v(\hat{Y}_d)}$$

$$= 9474.8 \pm 2\sqrt{1055724.7}$$

$$= 7419.8, 11529.8$$

$$\approx 7420, 11530$$

To summarize, the total number of pet animals is estimated as 9475, and it is almost sure that the actual number will fall in the closed interval [7420, 11530]. ■

It should be noted that the construction of reduced table 4.5, from the table 4.4, is not necessary. This has only been done to make the procedure easily understandable. One can keep selecting pairs of random numbers (i, j) using table 4.4, and if at any stage the random number i, first member of the pair (i, j), takes a value from among the serial numbers of the already selected population units, then such a random number should be discarded. This procedure will also yield a PPS without replacement sample.

4.8 MURTHY'S UNORDERED ESTIMATOR

The estimator considered in the preceding section depends on the order in which the units are drawn. There exists a more efficient estimator which ignores the order of selection of the units in the sample. Such an estimator is termed as *unordered estimator.* This estimator can be obtained by weighting all the possible ordered estimators with their respective probabilities. For instance, if a sample of two units, u_1 and u_2, is selected with varying probabilities WOR sampling, then Des Raj's ordered estimators $\hat{Y}_d(12)$ and $\hat{Y}_d(21)$ corresponding to the two possible orders of selections, (u_1, u_2) and (u_2, u_1), will be

$$\hat{Y}_d(12) = \frac{1}{2}\left[\frac{y_1}{p_1}(1+p_1) + \frac{y_2}{p_2}(1-p_1)\right] \tag{4.20}$$

$$\hat{Y}_d(21) = \frac{1}{2}\left[\frac{y_2}{p_2}(1+p_2) + \frac{y_1}{p_1}(1-p_2)\right] \tag{4.21}$$

and the probabilities of selecting these ordered samples are respectively

$$\phi(12) = \frac{p_1 p_2}{1-p_1} \tag{4.22}$$

$$\phi(21) = \frac{p_1 p_2}{1-p_2} \tag{4.23}$$

Under this set-up, the unordered estimator of total Y due to Murthy (1957) is then given by

$$\hat{Y}_m = \frac{\hat{Y}_d(12)\,\phi(12) + \hat{Y}_d(21)\,\phi(21)}{\phi(12) + \phi(21)}$$

On using (4.20), (4.21), (4.22), and (4.23), we get a simplified form of $\hat{Y}_m$ as in (4.24).

Murthy's unordered unbiased estimator of total Y for n = 2 :

$$\hat{Y}_m = \frac{1}{2-p_1-p_2}\left[\frac{y_1}{p_1}(1-p_2) + \frac{y_2}{p_2}(1-p_1)\right] \tag{4.24}$$

Variance of estimator $\hat{Y}_m$:

$$V(\hat{Y}_m) = \frac{1}{2}\sum_{i\neq j=1}^{N} \frac{P_i P_j (1-P_i-P_j)}{(2-P_i-P_j)}\left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right)^2 \tag{4.25}$$

Uniased estimator of variance $V(\hat{Y}_m)$:

$$v(\hat{Y}_m) = \frac{(1-p_1)(1-p_2)(1-p_1-p_2)}{(2-p_1-p_2)^2}\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2 \tag{4.26}$$

For sample of n units selected with PPS without replacement, Murthy's estimator $\hat{Y}_m$ becomes quite complicated, and its discussion is beyond the scope of this book.

Example 4.8

From the data given in example 4.7, and the sample of 2 villages selected there, estimate total number of pet animals in the area using Murthy's estimator. Also, construct the confidence interval for the population total.

Solution

The two villages, Jaspalon and Madpur, respectively at serial numbers 2 and 9 were included in the sample of size 2 drawn in example 4.7. From table 4.4, the y_i, p_i, and $(1-p_i)$ values corresponding to these two villages are

	y_i	p_i	$(1-p_i)$
Jaspalon :	690	.0657	.9343
Madpur :	680	.0819	.9181

Now from (4.24), we have Murthy's unordered estimator of total Y as

$$\hat{Y}_m = \frac{1}{2-p_1-p_2}\left[\frac{y_1}{p_1}(1-p_2) + \frac{y_2}{p_2}(1-p_1)\right]$$

$$= \frac{1}{(2-.0657-.0819)}\left[\frac{690}{.0657}(1-.0819) + \frac{680}{.0819}(1-.0657)\right]$$

$$= 9392.9$$

$$\approx 9393$$

Thus, the total number of pet animals in the 14 villages under consideration is estimated as 9393. The estimator of variance from (4.26) is computed as

$$v(\hat{Y}_m) = \frac{(1-p_1)(1-p_2)(1-p_1-p_2)}{(2-p_1-p_2)^2}\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2$$

$$= \frac{(1-.0657)(1-.0819)(1-.0657-.0819)}{(2-.0657-.0819)^2}\left(\frac{690}{.0657} - \frac{680}{.0819}\right)^2$$

$$= 1030832.7$$

Confidence limits for the total number of pet animals in the area, using Murthy's estimator, are then obtained as

$$\hat{Y}_m \pm 2\sqrt{v(\hat{Y}_m)}$$

$$= 9392.9 \pm 2\sqrt{1030832.7}$$

$$= 7362.3,\ 11423.5$$

$$\approx 7362,\ 11424$$

From the limits calculated above, one can be reasonably confident that the total number of pet animals in the area is in the range [7362, 11424]. ■

Order of selection of units in the sample was an important consideration in the development of Des Raj's and Murthy's estimators. Horvitz and Thompson (1952) suggested an estimator of population total which could be used with all kinds of without replacement samples. The estimator, however, requires the calculation of inclusion probabilities for the sample units. The estimator is efficient for sampling procedures, where inclusion probabilities for the population units are proportional to their sizes. Sampling procedures resulting in such probabilities are known as IPPS (*inclusion probability proportional to size)*, or πPS, designs. We now consider the estimator due to Horvitz and Thompson.

4.9 HORVITZ-THOMPSON ESTIMATOR

Horvitz and Thompson (1952) proposed a general estimator of population total which possesses several other very desirable properties (Sukhatme *et al.*, 1984). Yates and Grundy (1953) gave an elegant expression for the variance of this estimator. They, and Sen (1953), suggested an unbiased estimator of variance which reduces to zero when y_i is proportional to π_i, and assumes negative values less frequently as compared to the variance estimator proposed by Horvitz and Thompson (1952). Given below are the expressions for the Horvitz-Thompson (HT) estimator, its variance, and estimator of variance.

Horvitz-Thompson estimator of population total Y :

$$\hat{Y}_{ht} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \tag{4.27}$$

where π_i is the probability that i-th unit is included in the sample.

Variance of estimator $\hat{Y}_{ht}$:

$$V_{yg}(\hat{Y}_{ht}) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 \tag{4.28}$$

with π_{ij} denoting the probability of selecting pair of i-th and j-th units in the sample.

Unbiased estimator for variance $V_{yg}(\hat{Y}_{ht})$:

$$v_{yg}(\hat{Y}_{ht}) = \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{4.29}$$

Example 4.9

From the data given in example 4.7 and the sample selected there, estimate the total number of pet animals in the area using Horvitz-Thompson estimator. Also, build up the confidence interval for the population total.

Solution

In example 4.7, the villages bearing serial numbers 2 and 9 were included in a sample of size 2. The y_i, p_i, and $(1\text{-}p_i)$ values for the selected units, from table 4.4, are

	y_i	p_i	$(1\text{-}p_i)$
Jaspalon :	690	.0657	.9343
Madpur :	680	.0819	.9181

On using the values of p_i and $(1\text{-}p_i)$ in (4.13), (4.14), and (4.15), we find that

$$\sum_{i=1}^{14} \frac{P_i}{1-P_i} = 1.1466$$

$$\pi_1 = p_1\left(\sum_{i=1}^{14} \frac{P_i}{1-P_i} + 1 - \frac{p_1}{1-p_1}\right)$$

$$= .0657\,(1.1466+1-.0703)$$

$$= .1364$$

$$\pi_2 = p_2\left(\sum_{i=1}^{14} \frac{P_i}{1-P_i} + 1 - \frac{p_2}{1-p_2}\right)$$

$$= .0819\,(1.1466+1-.0892)$$

$$= .1685$$

$$\pi_{12} = p_1p_2\left(\frac{1}{1-p_1} + \frac{1}{1-p_2}\right)$$

$$= .0657\,(.0819)\left(\frac{1}{1-.0657} + \frac{1}{1-.0819}\right)$$

$$= .0116$$

Now HT estimator of total number of pet animals from (4.27) is computed as

$$\hat{Y}_{ht} = \frac{y_1}{\pi_1} + \frac{y_2}{\pi_2}$$

$$= \frac{690}{.1364} + \frac{680}{.1685}$$

$$= 9094.26$$

$$\approx 9094$$

Now we work out the estimate of variance $V(\hat{Y}_{ht})$. From (4.29), its expression for n=2 becomes

$$v_{yg}(\hat{Y}_{ht}) = \left(\frac{\pi_1\pi_2 - \pi_{12}}{\pi_{12}}\right)\left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2$$

The substitution of different values yields

$$v_{yg}(\hat{Y}_{ht}) = \left[\frac{(.1364)(.1685) - .0116}{.0116}\right]\left(\frac{690}{.1364} - \frac{680}{.1685}\right)^2$$

$$= 1027073.50$$

Also, the lower and upper limits of the confidence interval for the total number of pet animals, using HT estimator, are given by

$$\hat{Y}_{ht} \pm 2\sqrt{v_{yg}(\hat{Y}_{ht})}$$

$$= 9094.26 \pm 2\sqrt{1027073.50}$$

$$= 7067.37,\ 11121.15$$

$$\approx 7067,\ 11121$$

Hence, the total number of pet animals in the area under consideration is estimated as 9094. Its actual value is, however, expected to fall in the interval [7067, 11121]. ■

In the next two sections, we discuss two methods of selecting varying probabilities WOR samples that are comparatively easy to use in practice.

4.10 SEN-MIDZUNO METHOD

A simple procedure of selecting a PPS without replacement sample was suggested by Midzuno (1952), and also independently by Sen (1952). The steps involved in selecting a sample, and the corresponding inclusion probabilities for individual and pairs of units, are given below :

Steps for sample selection :

1. Select first unit using PPS without replacement.
2. Select (n-1) more units from the remaining (N-1) population units by SRS without replacement.

Inclusion Probabilities :

$$\pi_i = \frac{N-n}{N-1} P_i + \frac{n-1}{N-1} \tag{4.30}$$

$$\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(P_i + P_j) + \frac{n-2}{N-2}\right] \tag{4.31}$$

where i, j ($i \neq j$) = 1,2,...,N.

Under this scheme, the Yates-Grundy-Sen variance estimator $v_{yg}(\hat{Y}_{ht})$ in (4.29) would always be nonnegative. This is certainly encouraging, but the scheme is not entirely satisfactory in the sense that π_i is not proportional to P_i which is a minimum variance requirement. This requirement can be met if we select the sample using a new set of probabilities $\{P_i'\}$, known as *revised probabilities of selection*, where for i=1,2,...,N,

$$P_i' = nP_i \frac{N-1}{N-n} - \frac{n-1}{N-n}$$

Since P_i' is the probability, it must be nonnegative. This leads to the condition

$$P_i > \frac{n-1}{n(N-1)}$$

for all i=1, 2,..., N.

Thus, the use of revised probabilities $\{P_i'\}$, for getting efficient estimates of population total through Horvitz-Thompson estimator, will be possible in only those cases where the original selection probabilities satisfy the above condition. This condition on the initial probabilities usually does not hold, and hence limits the use of the scheme in practice. On using revised probabilities $\{P_i'\}$ in place of $\{P_i\}$, the inclusion probabilities in (4.30) and (4.31) reduce to those given below :

Inclusion probabilities using revised probabilities $\{P_i'\}$:

$$\pi_i = nP_i \tag{4.32}$$

$$\pi_{ij} = \frac{n(n-1)}{N-2}\left[P_i+P_j - \frac{1}{N-1}\right] \tag{4.33}$$

The Horvitz-Thompson estimator under the Sen-Midzuno scheme, with revised probabilities, is always more efficient than the usual estimator based on WR probability proportional to size sampling, for any sample size. The variance estimator $v_{yg}(\hat{Y}_{ht})$ under this scheme, with revised probabilities, is also always nonnegative.

Example 4.10

Extensive damage has been caused to wheat crop by hail storm in an area of Jagraon *tahsil* of Punjab state. The farmers of the affected area, consisting of 18 villages, approached the Government for compensation of this loss. In order to decide the amount of compensation, the administration needs to assess the total residual yield of the crop in the area. Keeping the objective in view, select a WOR sample of 4 villages using Sen-Midzuno method with initial selection probabilities proportional to the area under wheat crop. Estimate, therefrom, the total yield of wheat along with its standard error. Also, set up confidence interval for population total. The area under wheat crop (in hectares) for 18 villages of the population is given below in table 4.6. Though, the yield figures (in '000 quintals) are given for all the population villages for the sake of convenience, but in practice these will not be known in advance. These will be observed later for the units selected in the sample.

Solution

First we work out the initial selection probabilities $\{P_i = X_i/X\}$, noting that X, the total area under wheat crop, is 15715 hectares. The initial selection probabilities so calculated are also shown in the following table.

Table 4.6 Area under wheat and the initial probabilities

Village	Village name	Area under wheat (X_i)	Total yield (Y_i)	$P_i=X_i/X$
1	Akhara	920	21.160	.0585
2	Sujapur	250	4.550	.0159
3	Sidhwankhurd	220	4.312	.0140
4	Sohian	460	7.886	.0293
5	Sekhupura	235	5.464	.0150
6	Hanskalan	970	19.468	.0617
7	Kular	603	14.532	.0384
8	Kamalpur	785	13.785	.0500
9	Lakha	1425	30.068	.0907
10	Kaunke	2196	45.589	.1397
11	Gagrha	315	6.892	.0201
12	Dala	975	18.857	.0620
13	Dangia	426	8.375	.0271
14	Dholan	524	9.815	.0333
15	Mallah	1245	28.099	.0792
16	Bhammipur	1040	21.362	.0662
17	Manuke	1550	32.116	.0986
18	Rasulpur	1576	36.075	.1003
Total		X=15715	Y=328.405	1

In Sen-Midzuno method, the first unit is to be selected with probability proportional to the area under wheat and the remaining three by using equal probabilities WOR sampling. Using Lahiri's method of selection, let the selected pair of random numbers be (9,786). Since $786 \leq X_9$ (=1425), the village bearing serial number 9 is selected. As the selection procedure is WOR, the remaining units are now numbered 1 to 17. For selecting three more units by SRS without replacement, let the random numbers chosen between 1 and 17 be 7, 1, and 16. Thus the villages with these serial numbers, selected in the new set-up, are the 7th, 1st, and 17th villages in the old arrangement. Thus, the units included in the required sample are 1st, 7th, 9th, and 17th villages. The wheat production figures for these selected villages are observed to be 21.160, 14.532, 30.068, and 32.116 respectively. For estimating total yield, we shall use Horvitz-Thompson estimator. To do this, we need to calculate inclusion probabilities for the villages selected in the sample by making use of relations (4.30) and (4.31). Thus we have inclusion probability for the 1st sample unit, which is also the first population unit, as

$$\pi_1 = \frac{N-n}{N-1} P_1 + \frac{n-1}{N-1}$$

$$= \frac{(18-4)(.0585)}{(18-1)} + \frac{4-1}{18-1}$$

$$= .2246$$

Likewise, the inclusion probability for 2nd sample unit, which in this case happens to be 7th village in the population, is

$$\pi_2 = \frac{N-n}{N-1} P_7 + \frac{n-1}{N-1}$$

$$= \frac{(18-4)(.0384)}{(18-1)} + \frac{4-1}{18-1}$$

$$= .2081$$

Similarly, the inclusion probabilities for other sample units are

$$\pi_3 = \frac{N-n}{N-1} P_9 + \frac{n-1}{N-1}$$

$$= \frac{(18-4)(.0907)}{(18-1)} + \frac{4-1}{18-1}$$

$$= .2512$$

$$\pi_4 = \frac{N-n}{N-1} P_{17} + \frac{n-1}{N-1}$$

$$= \frac{(18-4)(.0986)}{(18-1)} + \frac{4-1}{18-1}$$

$$= .2577$$

Then, Horvitz-Thompson estimator for the population total, from (4.27), is given by

$$\hat{Y}_{ht} = \frac{y_1}{\pi_1} + \frac{y_2}{\pi_2} + \frac{y_3}{\pi_3} + \frac{y_4}{\pi_4}$$

$$= \frac{21.160}{.2246} + \frac{14.532}{.2081} + \frac{30.068}{.2512} + \frac{32.116}{.2577}$$

$$= 408.367$$

For computing Yates-Grundy-Sen estimator of the variance $V(\hat{Y}_{ht})$, we also need inclusion probabilities for pairs of units. Thus, following (4.31), the probability of selecting 1st and 7th population units together in the sample is

$$\pi_{12} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(P_1+P_7) + \frac{n-2}{N-2}\right]$$

$$= \frac{4-1}{18-1}\left[\frac{18-4}{18-2}(.0585 + .0384) + \frac{4-2}{18-2}\right]$$

$$= .0370$$

In the same way, the inclusion probability for 1st and 9th population units is

$$\pi_{13} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(P_1+P_9) + \frac{n-2}{N-2}\right]$$

$$= \frac{4-1}{18-1}\left[\frac{18-4}{18-2}(.0585 + .0907) + \frac{4-2}{18-2}\right]$$

$$= .0451$$

Similarly, the other inclusion probabilities for pairs of units are calculated as

$$\pi_{14} = \frac{4-1}{18-1}\left[\frac{18-4}{18-2}(.0585 + .0986) + \frac{4-2}{18-2}\right]$$

$$= .0463$$

$$\pi_{23} = \frac{4-1}{18-1}\left[\frac{18-4}{18-2}(.0384 + .0907) + \frac{4-2}{18-2}\right]$$

$$= .0420$$

$$\pi_{24} = \frac{4-1}{18-1}\left[\frac{18-4}{18-2}(.0384 + .0986) + \frac{4-2}{18-2}\right]$$

$$= .0432$$

$$\pi_{34} = \frac{4-1}{18-1}\left[\frac{18-4}{18-2}(.0907 + .0986) + \frac{4-2}{18-2}\right]$$

$$= .0513$$

Now from (4.29), Yates-Grundy-Sen estimator of variance is

$$v_{yg}(\hat{Y}_{ht}) = \frac{\pi_1\pi_2 - \pi_{12}}{\pi_{12}}\left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2 + \frac{\pi_1\pi_3 - \pi_{13}}{\pi_{13}}\left(\frac{y_1}{\pi_1} - \frac{y_3}{\pi_3}\right)^2$$

$$+ \frac{\pi_1\pi_4 - \pi_{14}}{\pi_{14}}\left(\frac{y_1}{\pi_1} - \frac{y_4}{\pi_4}\right)^2 + \frac{\pi_2\pi_3 - \pi_{23}}{\pi_{23}}\left(\frac{y_2}{\pi_2} - \frac{y_3}{\pi_3}\right)^2$$

$$+ \frac{\pi_2\pi_4 - \pi_{24}}{\pi_{24}}\left(\frac{y_2}{\pi_2} - \frac{y_4}{\pi_4}\right)^2 + \frac{\pi_3\pi_4 - \pi_{34}}{\pi_{34}}\left(\frac{y_3}{\pi_3} - \frac{y_4}{\pi_4}\right)^2$$

Substituting the values of π's in the above equation, one gets

$$
\begin{aligned}
v_{yg}(\hat{Y}_{ht}) &= \left[\frac{(.2246)(.2081) - .0370}{.0370}\right]\left(\frac{21.160}{.2246} - \frac{14.532}{.2081}\right)^2 \\
&+ \left[\frac{(.2246)(.2512) - .0451}{.0451}\right]\left(\frac{21.160}{.2246} - \frac{30.068}{.2512}\right)^2 \\
&+ \left[\frac{(.2246)(.2577) - .0463}{.0463}\right]\left(\frac{21.160}{.2246} - \frac{32.116}{.2577}\right)^2 \\
&+ \left[\frac{(.2081)(.2512) - .0420}{.0420}\right]\left(\frac{14.532}{.2081} - \frac{30.068}{.2512}\right)^2 \\
&+ \left[\frac{(.2081)(.2577) - .0432}{.0432}\right]\left(\frac{14.532}{.2081} - \frac{32.116}{.2577}\right)^2 \\
&+ \left[\frac{(.2512)(.2577) - .0513}{.0513}\right]\left(\frac{30.068}{.2512} - \frac{32.116}{.2577}\right)^2 \\
&= 1890.1699
\end{aligned}
$$

Estimate of standard error for the estimated total yield is

$$
\begin{aligned}
\text{se}(\hat{Y}_{ht}) &= \sqrt{1890.1699} \\
&= 43.476
\end{aligned}
$$

Following (2.8), the confidence interval for total yield is obtained from

$$
\begin{aligned}
&\hat{Y}_{ht} \pm 2\sqrt{v_{yg}(\hat{Y}_{ht})} \\
&= 408.367 \pm 86.952 \\
&= 321.415,\ 495.319
\end{aligned}
$$

To summarize, the estimated total residual wheat yield is 408367 quintals. However, its actual value is most likely to be in the range of 321415 to 495319 quintals. ■

4.11 RANDOM GROUP METHOD

This method has been proposed by Rao, Hartley, and Cochran (1962) and is commonly known as the *Rao, Hartley, Cochran* (RHC) *scheme*. Below are given the steps involved in selecting a sample of n units by using this method.

Steps involved in sample selection :

1. Split the population of size N units into n random groups of sizes $N_1, N_2, \ldots, N_n$ units, such that $\Sigma N_i = N$, i=1,2,..., n.
2. Select one unit with probability proportional to $\{P_i\}$ independently from each of these n groups.

The primary advantage of this scheme, in relation to the other WOR unequal probability sampling procedures, is that it does not involve heavy computations for drawing a sample even of size $n > 2$, and is, therefore, convenient to use in practice. Besides, the procedure does not require any restrictions on the initial probabilities of selection, and provides an unbiased estimator of variance which is always nonnegative and easy to compute.

If $(y_1, y_2,...,y_n)$ and $(p_1, p_2,...,p_n)$ are the values of the study variable and the initial selection probabilities for the selected sample units, then we have :

Unbiased estimator of population total Y :

$$\hat{Y}_{rhc} = \sum_{i=1}^{n} \frac{y_i}{p_i} \phi_i \tag{4.34}$$

where ϕ_i is the sum of initial selection probabilities for all the N_i units in the i-th random group.

Variance of the estimator $\hat{Y}_{rhc}$:

$$V(\hat{Y}_{rhc}) = \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)} \sum_{i=1}^{N} \left(\frac{Y_i}{P_i} - Y\right)^2 P_i \tag{4.35}$$

Estimator of variance $V(\hat{Y}_{rhc})$:

$$v(\hat{Y}_{rhc}) = \frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{Y}_{rhc}\right)^2 \phi_i \tag{4.36}$$

If the random groups are of equal sizes, so that $N_i = N/n$, the expressions for the variance $V(\hat{Y}_{rhc})$ and its estimator $v(\hat{Y}_{rhc})$ reduce to simpler forms.

For random groups of equal sizes :

$$V(\hat{Y}_{rhc}) = \frac{N-n}{n(N-1)} \sum_{i=1}^{N} \left(\frac{Y_i}{P_i} - Y\right)^2 P_i \tag{4.37}$$

$$v(\hat{Y}_{rhc}) = \frac{N-n}{N(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{Y}_{rhc}\right)^2 \phi_i \tag{4.38}$$

The efficiency of the estimator $\hat{Y}_{rhc}$ is maximum when random groups are of equal size. This could be possible only in cases where the population size N is completely divisible by n. The investigator should, therefore, keep the group sizes as equal as

possible. Thus if N = nQ+R, where Q is the quotient in division of N by n and R is the remainder, then the investigator should keep the sizes of the first (n-R) groups as Q, whereas the sizes of remaining R groups be taken as (Q+1).

Example 4.11

From the data given in example 4.10, estimate the total residual yield of wheat by selecting a WOR sample of size 4, with probability proportional to area under wheat crop, through RHC scheme. Also, estimate the standard error of your estimate, and construct confidence interval for the population total.

Solution

Since we are to select a sample of size 4, four random groups need to be constructed. Thus, first two groups are formed of 4 units and the remaining two groups consist of 5 units. For this, we choose 18 random numbers from 1 to 18, without replacement, which amounts to arranging the given population in a random order. Let the random numbers (discarding repetitions) from 1 to 18, thus chosen, be 1, 12, 11, 6, 8, 2, 17, 13, 18, 10, 4, 15, 5, 9, 3, 7, 14, and 16. The villages bearing serial numbers corresponding to the first four selected random numbers constitute the Ist group, whereas the next four form group II. Five villages corresponding to the next five random numbers constitute group III, and the remaining five villages are included in group IV. Thus, we have table 4.7 in which

Y_{ij} = total yield, in thousand quintals, for the j-th village in the i-th random group,
X_{ij} = area under wheat crop, in hectares, for the j-th village in the i-th random group, and
P_{ij} = initial probability of selection for the j-th village in the i-th random group.

Table 4.7 Random group formation of 18 population units

Group I				Group II			
Serial No.	Y_{1j}	X_{1j}	P_{1j}	Serial No.	Y_{2j}	X_{2j}	P_{2j}
1	21.160	920	.0585	1	13.785	785	.0500
2	18.857	975	.0620	2	4.550	250	.0159
3	6.892	315	.0201	3	32.116	1550	.0986
4	19.468	970	.0617	4	8.375	426	.0271
Total		3180	.2023	Total		3011	.1916

Group III				Group IV			
Serial No.	Y_{3j}	X_{3j}	P_{3j}	Serial No.	Y_{4j}	X_{4j}	P_{4j}
1	36.075	1576	.1003	1	30.068	1425	.0907
2	45.589	2196	.1397	2	4.312	220	.0140
3	7.886	460	.0293	3	14.532	603	.0384
4	28.099	1245	.0792	4	9.815	524	.0333
5	5.464	235	.0150	5	21.362	1040	.0662
Total		5712	.3635	Total		3812	.2426

Using Lahiri's method of selection in each group, and referring to the table of random numbers, the admissible pairs of random numbers (excluding the rejected pairs of random numbers) selected, in this case, for groups I, II, III, and IV respectively are (4, 556), (4, 286), (1, 946), and (5, 868). The population units selected in the sample are thus with serial numbers 6, 13, 18, and 16 in table 4.6. These units respectively have y-values as 19.468, 8.375, 36.075, and 21.362.

Now the RHC estimator of total Y is given by

$$\hat{Y}_{rhc} = \frac{y_1}{p_1}\phi_1 + \frac{y_2}{p_2}\phi_2 + \frac{y_3}{p_3}\phi_3 + \frac{y_4}{p_4}\phi_4$$

On substituting these values from table 4.7, one gets the value of the estimate as

$$\hat{Y}_{rhc} = \frac{(19.468)(.2023)}{.0617} + \frac{(8.375)(.1916)}{.0271} + \frac{(36.075)(.3635)}{.1003} + \frac{(21.362)(.2426)}{.0662}$$

$$= 332.068$$

Therefore, the estimated total residual yield of wheat for all the 18 villages is 332068 quintals.

From (4.36), we have the estimator of $V(\hat{Y}_{rhc})$ as

$$v(\hat{Y}_{rhc}) = \left[\frac{N_1^2 + N_2^2 + N_3^2 + N_4^2 - N}{N^2 - (N_1^2 + N_2^2 + N_3^2 + N_4^2)}\right]\left[\left(\frac{y_1}{p_1} - \hat{Y}_{rhc}\right)^2 \phi_1 + \left(\frac{y_2}{p_2} - \hat{Y}_{rhc}\right)^2 \phi_2 + \left(\frac{y_3}{p_3} - \hat{Y}_{rhc}\right)^2 \phi_3 + \left(\frac{y_4}{p_4} - \hat{Y}_{rhc}\right)^2 \phi_4\right]$$

On making substitutions, one gets

$$v(\hat{Y}_{rhc}) = \left[\frac{4^2 + 4^2 + 5^2 + 5^2 - 18}{18^2 - (4^2 + 4^2 + 5^2 + 5^2)}\right]\left[\left(\frac{19.468}{.0617} - 332.068\right)^2 (.2023)\right.$$

$$+\left(\frac{8.375}{.0271} - 332.068\right)^2 (.1916) + \left(\frac{36.075}{.1003} - 332.068\right)^2 (.3635)$$

$$+\left(\frac{21.362}{.0662} - 332.068\right)^2 (.2426)\Bigg]$$

$$= 120.397$$

The estimated standard error for the estimate of population total will simply be square root of $v(\hat{Y}_{rhc})$. This gives

$$se(\hat{Y}_{rhc}) = \sqrt{120.397}$$

$$= 10.973$$

Confidence interval for Y, following (2.8), can be obtained from

$$\hat{Y}_{rhc} \pm 2\, se(\hat{Y}_{rhc})$$

$$= 332.068 \pm 2(10.973)$$

$$= 310.122,\ 354.014$$

The total residual yield of wheat crop for the area under study is, therefore, likely to be in the range of 310122 to 354014 quintals. ■

Since RHC method is a commonly used varying probability WOR sampling procedure, we now look into its efficiency aspect.

4.12 RELATIVE EFFICIENCY OF RHC ESTIMATOR

Here we consider relative efficiency of RHC estimator with respect to (1) PPS with replacement, and (2) SRS without replacement.

4.12.1 *In Relation to PPSWR Sampling*

From (4.2) and (4.35), one can see that the relative efficiency of RHC strategy with respect to PPS with replacement method is given by

$$RE = \frac{V(\hat{Y}_{pps})}{V(\hat{Y}_{rhc})}$$

The above equation reduces to

$$\left.\begin{aligned} RE &= \frac{N(N-1)}{n\left(\sum_{i=1}^{n} N_i^2 - N\right)} && \text{(for unequal size groups)} \\ &= \frac{N-1}{N-n} && \text{(for groups of equal size)} \end{aligned}\right\} \quad (4.39)$$

Thus, when random groups are of equal size, the relative efficiency will always be more than 1. Further, the RE defined above does not depend on the population variability. The estimates of variances $V(\hat{Y}_{pps})$ and $V(\hat{Y}_{rhc})$ are, therefore, not required for obtaining estimated relative efficiency from the selected sample. Hence, in practice, once the sizes of random groups and the sample size are known, the actual relative efficiency can be determined.

4.12.2 *In Relation to SRS Without Replacement*

Also, the estimator $\hat{Y}_{rhc}$ is expected to fare well in relation to the usual SRS without replacement estimator, when the auxiliary variable is approximately proportional to estimation variable y, and the line of regression passes through (or nearly through) the origin. The percent relative efficiency of estimator $\hat{Y}_{rhc}$ with respect to usual SRS without replacement estimator $\hat{Y}$ will be defined as

$$RE = \frac{V(\hat{Y})}{V(\hat{Y}_{rhc})}\,(100) \tag{4.40}$$

where $V(\hat{Y})$ and $V(\hat{Y}_{rhc})$ are given in (3.12) and (4.35) respectively. In order to calculate relative efficiency defined in (4.40), for a particular population, one may proceed as in example 4.4. However, in actual practice, the population data are not available, and the investigator has to estimate relative efficiency from the sample observations. For calculating estimated relative efficiency, one needs estimates of involved variances.

An unbiased variance estimator for the estimator $\hat{Y}$ on the basis of a PPS sample, selected by RHC scheme, is given in (4.41).

$$v_{rhc}(\hat{Y}) = \frac{N-n}{n(N-1)}\left[N\sum_{i=1}^{n}\frac{y_i^2}{p_i}\phi_i - \hat{Y}_{rhc}^2 + v(\hat{Y}_{rhc})\right] \tag{4.41}$$

Thus, the percent relative efficiency, defined in (4.40), is estimated by

$$RE = \frac{v_{rhc}(\hat{Y})}{v(\hat{Y}_{rhc})}\,(100)$$

For illustration, we consider an example.

Example 4.12

From the sample data of example 4.11, estimate percent relative efficiency of RHC estimator in relation to without replacement SRS based estimator $\hat{Y}$

Solution

From example 4.11, we have N=18, n=4, $\hat{Y}_{rhc}$=332.068, and $v(\hat{Y}_{rhc})$=120.397. The values of y for the units selected in example 4.11, along with their original probabilities (from table 4.6) and ϕ_i's (from table 4.7), are

Unit No.	:	6	13	18	16
y_i	:	19.468	8.375	36.075	21.362
p_i	:	.0617	.0271	.1003	.0662
ϕ_i	:	.2023	.1916	.3635	.2426

From these values, we obtain

$$\sum_{i=1}^{4} \frac{y_i^2}{p_i}\phi_i = \frac{(19.468)^2\,(.2023)}{.0617} + \frac{(8.375)^2\,(.1916)}{.0271} + \frac{(36.075)^2\,(.3635)}{.1003} + \frac{(21.362)^2\,(.2426)}{.0662}$$

$$= 8127.335$$

On substituting in (4.41) the values of different terms, we get the value of estimator $v_{rhc}(\hat{Y})$. Therefore,

$$v_{rhc}(\hat{Y}) = \frac{18-4}{4\,(18-1)}\,[(18)\,(8127.335) - (332.068)^2 + 120.397]$$

$$= 7441.262$$

The required percent relative efficiency will, therefore, be

$$RE = \frac{7441.262}{120.397}\,(100)$$

$$= 6180.60\ \blacksquare$$

4.13 SOME FURTHER REMARKS

4.1 Goodman and Kish (1950) suggested selection of a systematic sample with varying probabilities, while Sampford (1967) proposed a method of selecting a PPS without replacement sample of two units. There are many other IPPS sampling procedures given by Narain (1951), Madow (1949), Brewer (1963), Rao (1965), Durbin (1967), and Hanurav (1967). Use of these methods in practice is, however, cumbersome. Recently, Dey and Srivastava (1987) have proposed an IPPS without replacement sampling strategy which is relatively easy to use in practice for any sample size.

4.2 Recently, Mangat (1993) has proposed an alternative estimator for RHC scheme. Though, his estimator is more efficient than RHC estimator $\hat{Y}_{rhc}$ for certain situations, but the expressions for its variance and variance estimator are complicated.

4.3 The relative efficiency of the RHC estimator with respect to PPS with replacement sampling estimator, for a given sample size, was discussed in section 4.12. Since there is a possibility of some population units getting selected more than once in PPS with replacement sampling, the effective sample size for this

sampling scheme will be less than that in RHC scheme where the sample units are selected without replacement. Therefore, the expected survey cost for the former scheme will be less than the latter. Singh and Kishore (1975) have shown that for a fixed total cost of the survey, PPS with replacement estimator can sometimes be more efficient than the RHC estimator.

4.4 An interesting modification of RHC strategy is due to Singh and Lal (1978). They have proposed construction of first random group in RHC method using Sen-Midzuno scheme, whereas remaining groups are constructed in the usual manner.

4.5 The estimators discussed in this chapter are suitable for characters which are highly and positively correlated with the selection probabilities. In multicharacter surveys, some of the study variables may be positively and highly correlated with the selection probabilities, while others could be poorly or even negatively correlated. The estimators due to Rao (1966) and Bansal and Singh (1985) would be found more appropriate in such situations.

4.6 When the regression line of y on x meets the y-axis away from origin, or the regression is not linear, the use of the estimator in (4.1), as such, is not recommended. If the regression of y on x is of the form $y=\phi(x)$, where $\phi(x)$ is a real valued function of x and $\phi(x)$ is taken as the new size variable z, the regression of y on z will then be linear through origin. An estimator of population total based on this concept has been suggested by Singh and Gupta (1972).

LET US DO

4.1 Which situations warrant the use of varying probability sampling in place of simple random sampling ?

4.2 What are the various steps involved in cumulative total method of selecting a with replacement PPS sample? Why does it become little difficult to use the method in case of large populations ?

4.3 Describe Lahiri's method for selecting a varying probabilities WR sample. Why, and when, is this method preferred over cumulative total method?

4.4 Give the estimator of population total, and the estimator of its variance, for a PPS with replacement sample. How will you modify it to get corresponding estimators for population mean ?

4.5 A television team is interested in estimating total number of votes cast up to 1 p.m. on the day of Assembly elections in a rural constituency. The total number of villages in the constituency is 36, and the available frame showing total number of persons eligible to vote in each village is given in the following table.

Village	Total votes	Village	Total votes	Village	Total votes	Village	Total votes
1	900	10	576	19	341	28	681
2	1340	11	1083	20	649	29	990
3	860	12	1644	21	1366	30	1232
4	1716	13	871	22	1199	31	749
5	405	14	605	23	890	32	836
6	704	15	970	24	667	33	910
7	816	16	1413	25	1380	34	1060
8	1426	17	1136	26	571	35	1270
9	1113	18	870	27	1570	36	1710

Select a PPS with replacement sample of 6 villages, taking size variable as the total number of votes in the village, using (1) cumulative total method, and (2) Lahiri's method.

4.6 For the job outlined in exercise 4.5, suppose the six villages bearing serial numbers 35, 36, 12, 7, 36, and 27 were included in the sample. The number of votes cast up to 1 p.m. for the selected villages are given below :

Village	Total votes	Votes cast	Village	Total votes	Votes cast
35	1270	578	7	816	517
36	1710	780	36	1710	780
12	1644	698	27	1570	1121

Estimate total number of votes cast up to 1 p.m. in the constituency, and place confidence limits on it.

4.7 Estimate the relative efficiency of the estimator $\hat{Y}_{pps}$ with respect to the usual with replacement SRS estimator from the PPS sample selected in exercise 4.6.

4.8 Describe the rule for determining sample size, for a given precision of the estimator, in case of varying probabilities WR sampling.

4.9 Assume that the sample of 6 villages drawn in exercise 4.6 is a preliminary sample. Using this sample information, determine the sample size required to estimate the total number of votes cast up to 1 p.m. in all the36 villages of the constituency with a margin of error not exceeding 2000 votes. Take $\alpha = .05$.

4.10 Discuss, why is it more difficult to select a varying probabilities WOR sample in comparison to a PPS with replacement sample?

4.11 Differentiate between Des Raj's ordered estimator and Murthy's unordered estimator of population total.

4.12 For the population of exercise 4.5, suppose that the 4th and 33rd villages were selected to constitute a without replacement PPS sample, where the initial selection

probabilities are proportional to the total number of votes. The information corresponding to these villages was obtained as

Village	:	4	33
Total votes	:	1716	910
Votes cast	:	810	613

Estimate the total number of votes cast up to 1 p.m. using (1) Des Raj's ordered estimator, and (2) Murthy's unordered estimator. Also, estimate the variances of the two estimates, and build up the confidence interval in each case.

4.13 What is Horvitz-Thompson estimator ? Discuss its merits and demerits.

4.14 Describe Sen-Midzuno's method for selecting a PPS without replacement sample. Discuss, why should the original probabilities of selection in this method satisfy a condition to yield inclusion probabilities, for various population units, to be proportional to their sizes ?

4.15 Sen-Midzuno method of selecting a PPS without replacement sample is quite easy to use in practice but still it is not very popular with the statisticians. Why is it so ?

4.16 There are 12 higher secondary schools in a development block. The total number of teachers in each of these 12 schools are as under :

School	:	1	2	3	4	5	6	7	8	9	10	11	12
Teachers	:	21	13	16	12	19	11	14	17	13	16	15	27

The Block Education Officer wishes to have a quick estimate of the total number of teachers in these schools, who reach their schools late by half an hour or more. It was decided to conduct a small survey for getting this estimate. Taking total number of teachers in the school as the size variable, the schools at serial numbers 1, 3, and 12 were included in the sample, using Sen-Midzuno scheme. On a randomly selected day, 6, 4, and 8 teachers were found reaching late in schools at serial numbers 1, 3, and 12 respectively. Estimate the total number of teachers coming late in all the 12 higher secondary schools on that day. Also, work out the confidence interval for population value.

4.17 Explain the steps involved in RHC method of selecting a PPS without replacement sample. In what sense, this procedure is more desirable than the other procedures discussed in the chapter ?

4.18 The data in respect of geographical area (in hectares) and the number of coconut trees are given below for 12 villages of Orissa state in India.

Village	Area	Trees	Village	Area	Trees
1	737	1700	7	492	900
2	911	2163	8	563	972
3	603	1814	9	864	1407
4	815	1966	10	971	1516
5	403	817	11	578	706
6	379	763	12	484	617

Select a sample of 3 villages with probability proportional to total geographical area using RHC scheme. From the selected units, estimate the total number of coconut trees in the 12 villages, and place confidence limits on the population value.

4.19 Given below are four hypothetical population values $\{Y_i\}$, considered by Yates and Grundy (1953), along with their original probabilities $\{P_i\}$.

P_i	:	.1	.2	.3	.4
Y_i	:	.5	1.2	2.1	3.2

For n=2, work out the sampling variances of the following estimators of population total :

a. PPS with replacement estimator
b. Des Raj's ordered estimator
c. Murthy's unordered estimator
d. Horvitz-Thompson estimator
e. Estimator based on Sen-Midzuno scheme
f. RHC estimator

4.20 Work out percent relative efficiency of RHC estimator $\hat{Y}_{rhc}$ over PPS with replacement estimator $\hat{Y}_{pps}$, if the sample of size 9 is to be drawn from a population consisting of 72 units.

4.21 Is the PPS sampling always more efficient than equal probability sampling? Comment.

4.22 There are 15 fish catching centers along a stretch of Indian coast. The total number of boats landing in a day (x) and the corresponding catch of fish in quintals (y) is given below :

Center	y	x	Center	y	x	Center	y	x
1	8.20	40	6	10.00	51	11	7.08	37
2	2.60	30	7	12.80	68	12	7.51	39
3	8.75	45	8	6.70	33	13	8.80	47
4	11.70	60	9	8.50	48	14	7.80	40
5	11.00	55	10	9.30	54	15	9.00	44

Work out the relative efficiency of estimator $\hat{Y}_{rhc}$ of the total catch of fish with respect to the usual WOR simple random sample estimator $\hat{Y}$ for a sample of 5 centers. Take initial selection probabilities proportional to the number of boats landing in a day.

4.23 Estimate the percent relative efficiency of RHC estimator over the SRS without replacement estimator from the PPS sample selected in exercise 4.18.

CHAPTER 5

Stratified Sampling

5.1 INTRODUCTION

The precision of an estimate of the population mean or total, besides sample size, also depends on the variability among the units of the population. Therefore, apart from increasing the sample size, another possible way to increase the precision of the estimate could be to divide the population units into certain number of groups, such that the variability within the groups is minimum whereas it is maximum between the groups. Smaller samples could then be selected from each of the groups so formed, such that the total number of sampled units over all the groups equal the required overall sample size. The groups thus formed are called *strata*, and the process of forming strata is known as *stratification*. In this connection, we have the following definitions :

Definition 5.1 The procedure of partitioning the population into groups, called strata, and then drawing a sample independently from each stratum, is known as *stratified sampling.*

Definition 5.2 If the sample drawn from each stratum is random one, the procedure is then termed as *stratified random sampling.*

In case of stratified simple random sampling, since the samples from different strata are selected independently, each stratum can, therefore, be treated as a separate population. All the results given in chapter 3 can thus be applied to each stratum.

The stratified mean estimator will be more efficient than the usual simple random sample mean if variation between the strata means is sufficiently large in relation to within stratum variation. The extent of gain in precision, however, also depends on the method used for selecting the units from each stratum. Once the procedure of selecting units from the strata is finalized, the other points that need careful consideration are :

1. determining the number of strata to be constructed,
2. allocation of total sample size to different strata, and
3. the choice of strata.

The answers to the above points are to be such that they minimize sampling variance for a given cost, or the cost is minimized for a specified precision. The exact solutions to these problems depend on the values of study variable (also called *estimation variable)* for all population units, which are never available. Hence, the solutions are to be based on the similar data available for some suitable supplementary variable (called *stratification*

variable when strata are constructed on it), and on the knowledge regarding the relationship between this variable and the estimation variable.

Given below are some broad *principles* that should be kept in mind, while going for stratified sampling.

1. The strata should be nonoverlapping, and should together comprise the whole population.
2. The units forming any stratum should be similar with respect to the study variable, so that, the variability within each stratum is reduced.
3. When it is difficult to stratify the population with respect to study variable, or a highly correlated auxiliary variable, the administrative convenience may be considered as the basis for stratification . However, the gain in precision can not be guaranteed in this case, since the stratification chosen purely for administrative convenience will not necessarily yield the relative homogeneity within the strata.

We now point out some of the *advantages*, that the stratified random sampling enjoys over unstratified sampling. These are briefly discussed below :

1. Since the population is first divided into various strata, and then samples are drawn from each stratum, there is little possibility of any essential group of population being completely excluded. Hence, stratification ensures that a better cross section of the population is represented in the sample as compared to that under unstratified sampling.
2. The stratification makes it possible to use different sampling designs in different strata thereby enabling effective utilization of the available auxiliary information. It is particularly true in cases, where the extent and nature of the available information vary from stratum to stratum. Separate estimates obtained for different strata can be combined into a precise estimate for the whole population.
3. When a survey organization has field offices in several zones, it might be desirable to treat the zones as strata from the point of view of administrative convenience, as it will facilitate the supervision and organization of field work.
4. When there are extreme values in population, these can be grouped into a separate stratum thereby reducing the variability within other strata.
5. The geographical and topographical considerations may also be the reason for resorting to stratification. There may be different types of sampling problems in plains, deserts, and hilly areas. These may need different approaches for their resolution. Hence, it would be advantageous to form separate stratum for each of such areas.
6. Since the variability within strata is considerably reduced, the stratification normally provides more efficient estimates than the usual unstratified sampling.
7. The cost of conducting the survey is expected to be less for stratified sampling, when strata are formed keeping administrative convenience in mind.

From the foregoing discussion, we thus conclude that the stratification might be an aid to efficient estimation. It is, therefore, worth considering the procedure in greater detail.

5.2 NOTATIONS

Unless specified otherwise, throughout this chapter, we shall assume the sampling within each stratum to be simple random sampling WOR. The suffix h stands for h-th stratum, h=1,2,...,L, where L denotes the total number of strata into which the population has been divided. Similarly, the suffix i will indicate the i-th unit within the stratum. All the following symbols refer to the h-th stratum :

N_h = total number of units in the stratum

n_h = number of units selected in the sample from the stratum

$W_h = N_h/N$ = proportion of the population units falling in the stratum

$f_h = n_h/N_h$ = sampling fraction for the stratum

Y_{hi} = the value of study variable for the i-th unit in the stratum, i=1,2,...,N_h

$Y_h = \sum_{i=1}^{N_h} Y_{hi}$ = stratum total for the estimation variable based on N_h units

$\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$ = mean for the estimation variable in the stratum

$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ = stratum sample mean for the estimation variable

$\sigma_h^2 = \frac{1}{N_h} (\sum_{i=1}^{N_h} Y_{hi}^2 - N_h \bar{Y}_h^2)$ = stratum variance based on N_h units

$S_h^2 = \frac{N_h}{N_h - 1} \sigma_h^2$ = stratum mean square based on N_h units

$s_h^2 = \frac{1}{n_h - 1} (\sum_{i=1}^{n_h} y_{hi}^2 - n_h \bar{y}_h^2)$ = sample mean square based on n_h sample units drawn from the stratum

We now consider the problem of estimating population mean or total from a stratified simple random sample.

5.3 ESTIMATION OF MEAN AND TOTAL USING SIMPLE RANDOM SAMPLING

In stratified sampling, it is important to keep in mind that the samples are drawn independently from each stratum, so that, the strata estimates are not correlated. From chapter 3, we know that the sample mean $\bar{y}_h$ is an unbiased estimator of the stratum mean $\bar{Y}_h$, which implies that $N_h\bar{y}_h$ is an unbiased estimator of stratum total $N_h\bar{Y}_h$. It is, therefore,

reasonable to arrive at the following estimator of population mean $\bar{Y}$. We denote this estimator by $\bar{y}_{st}$, where the subscript (st) stands for stratified.

5.3.1 *Stratified SRS Without Replacement*

The unbiased estimator of population mean and the other related expressions for this case are listed below :

Unbiased estimator of population mean :

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{5.1}$$

Variance of estimator $\bar{y}_{st}$:

$$\begin{aligned} V(\bar{y}_{st}) &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) S_h^2 \\ &= \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \\ &= \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h - 1}{N_h - 1}\right) \frac{\sigma_h^2}{n_h} \end{aligned} \tag{5.2}$$

Estimator of variance $V(\bar{y}_{st})$:

$$v(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) s_h^2 \tag{5.3}$$

5.3.2 *Stratified SRS With Replacement*

If the sample from each stratum is selected by SRS with replacement, the expressions for variance and estimator of variance for the stratified estimator of population mean follow from relations (5.2) and (5.3), as explained in section 3.3.2, by taking fpc = $[1- (n_h-1)/(N_h-1)]$ equal to one and the sampling fraction $f_h = n_h/N_h$ as zero, for h = 1, 2,...,L. Thus we have :

Variance of estimator $\bar{y}_{st}$:

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 \sigma_h^2}{n_h} \tag{5.4}$$

Estimator of variance $V(\bar{y}_{st})$:

$$v(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h} \tag{5.5}$$

As mentioned in the earlier chapters, the estimator $\hat{Y}_{st}$ of population total Y can be obtained by multiplying the estimator of mean $\bar{y}_{st}$ by N. Also, the variance and estimator of variance expressions for $\hat{Y}_{st}$ are obtained by multiplying $V(\bar{y}_{st})$ and $v(\bar{y}_{st})$ respectively by N^2, using the expressions of $V(\bar{y}_{st})$ and $v(\bar{y}_{st})$ for WOR or WR sampling as the case may be. Accordingly, the lower and upper confidence limits for the population mean are to be multiplied by N to yield the corresponding expressions for the total Y.

Example 5.1

An assignment was given to four students attending a sample survey course. The problem was to estimate the average time per week devoted to study in Punjab Agricultural University (PAU) library by the students of this university. The university is running undergraduate, master's degree and doctoral programs. Number of students registered for the three programs is 1300, 450, and 250 respectively. Since the value of the study variable is likely to differ considerably with the program, the investigator divided the population of students into 3 strata: undergraduate program (stratum I), master's program (stratum II), and doctoral program (stratum III). First of the four students selected WOR simple random samples of sizes 20, 10, and 12 students from strata I, II, and III respectively, so that, the total sample is of size 42. The information about weekly time devoted in library is given in table 5.1.

Table 5.1 Time (in hours) devoted to study in the university library during a week

Stratum I			Stratum II		Stratum III	
0	1	9	12	6	10	24
4	4	4	9	10	14	15
3	3	6	11	9	20	14
5	6	1	13	11	11	18
2	8	2	8	7	16	19
0	10	3			13	20
3	2					

Estimate the average time per week devoted to study by a student in PAU library. Also, build up the confidence interval for this average.

Solution

Proceeding with the solution, we first prepare table 5.2 presenting calculated values of strata sample means and sample mean squares.

Table 5.2 Calculated values of strata weights, sample means, and sample mean squares

Stratum I	Stratum II	Stratum III
$n_1 = 20$	$n_2 = 10$	$n_3 = 12$
$N_1 = 1300$	$N_2 = 450$	$N_3 = 250$
$W_1 = .650$	$W_2 = .225$	$W_3 = .125$
$\bar{y}_1 = 3.800$	$\bar{y}_2 = 9.600$	$\bar{y}_3 = 16.167$
$s_1^2 = 7.958$	$s_2^2 = 4.933$	$s_3^2 = 17.049$

The stratum weight W_h, sample mean $\bar{y}_h$, and sample mean square s_h^2 have been defined earlier in section 5.2. For calculation of $\bar{y}_h$ and s_h^2, one is to proceed in the same way as for $\bar{y}$ and s^2 in chapter 3. Now, the estimate of average time (in hours) per week devoted to study by a student in the university library, is

$$\begin{aligned}\bar{y}_{st} &= \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3) \\ &= \frac{1}{2000}[1300\,(3.800) + 450\,(9.600) + 250\,(16.167)] \\ &= 6.651\end{aligned}$$

Also, the estimate of variance is computed from (5.3) as

$$\begin{aligned}v(\bar{y}_{st}) &= \frac{W_1^2(N_1 - n_1)\,s_1^2}{N_1 n_1} + \frac{W_2^2\,(N_2 - n_2)\,s_2^2}{N_2 n_2} + \frac{W_3^2(N_3 - n_3)\,s_3^2}{N_3 n_3} \\ &= \frac{(.650)^2\,(1300 - 20)\,(7.958)}{(1300)\,(20)} + \frac{(.225)^2\,(450 - 10)\,(4.933)}{(450)\,(10)} \\ &\quad + \frac{(.125)^2\,(250 - 12)\,(17.049)}{(250)\,(12)} \\ &= .16553 + .02442 + .02113 \\ &= .21108\end{aligned}$$

Using (2.8), we obtain the limits of confidence interval as

$$\begin{aligned}&\bar{y}_{st} \pm 2\sqrt{v(\bar{y}_{st})} \\ &= 6.651 \pm 2\sqrt{.21108} \\ &= 5.732,\ 7.570\end{aligned}$$

Thus, the average time per week devoted to study by a student in PAU library, falls in the closed interval [5.732, 7.570] hours, with probability approximately equal to .95. ■

5.4 ALLOCATION OF SAMPLE SIZE

Although the total sample size n is generally limited by the budget available for a survey, the allocation of the total sample to the strata remains at the discretion of the investigator. The precision of the estimator of population mean based on stratified sample also depends on the allocation of sample to different strata. Arbitrary allocation of the overall sample to different strata, as considered in example 5.1, is not based on any criterion, and hence does not seem reasonable. Intuitively, one may feel that the principal factors that should be kept in mind in this case are the stratum size, variability within stratum, and the cost of taking observations per sampling unit in the stratum. So far as the cost aspect is concerned, we consider one of the simplest cost functions as

$$C = c_o + \sum_{h=1}^{L} c_h n_h \tag{5.6}$$

where c_o is the overhead cost, which includes the cost of designing the questionnaire, selection of the sample, and analysis of survey data, etc. Also, c_h is the cost of observing study variable y for each unit selected in the sample from h-th stratum, h=1, 2,...,L.

A good allocation is one, where maximum precision is obtained with minimum resources. Various workers have proposed different methods to achieve this aim. However, we shall discuss only the commonly used methods of sample allocation.

Methods of sample allocation to different strata :

1. Equal allocation
2. Proportional allocation
3. Optimum allocation

We now briefly discuss these methods of sample allocation.

5.4.1 *Equal Allocation*

In case of *equal allocation*, number of sampling units selected from each stratum is equal. Thus for h = 1, 2,...,L,

$$n_h = \frac{n}{L}$$

units will be selected from each stratum. On substituting the above value of n_h in the cost constraint (5.6), one gets the required total sample size. This would be

$$n = \frac{L\,(C - c_0)}{\sum_{h=1}^{L} c_h}$$

Sample size for h-th stratum in case of equal allocation :

$$n_h = \frac{n}{L} \tag{5.7}$$

Total sample size for fixed total cost :

$$n = \frac{L\,(C - c_0)}{\sum_{h=1}^{L} c_h} \tag{5.8}$$

On substituting the value of $n_h = n/L$ in expressions from (5.2) to (5.5) for variance and estimator of variance, one may get the expressions appropriate for equal allocation.

It may be pointed out here, that this method of sample allocation is used when strata sizes do not differ much from each other, and the information about the variation within the strata is lacking.

Example 5.2

Second student in the group of four, was asked to independently take up the estimation problem given in example 5.1, using equal allocation. He was provided with \$150, including overhead cost of \$ 24. The cost of contacting the students, and collecting information is \$ 3 per student. How many students would he select in the sample, for collecting the desired information ?

Solution

The given details are: N_1=1300, N_2 = 450, N_3 = 250, L = 3, C = \$150, c_0 = \$24, and $c_1 = c_2 = c_3$ = \$3. The total number of students that could be included in sample is given by (5.8). Thus,

$$n = \frac{L\,(C - c_0)}{\sum_{h=1}^{L} c_h}$$

$$= \frac{3\,(150 - 24)}{3+3+3}$$

$$= 42 \blacksquare$$

Example 5.3

In example 5.2, the second student from the group of four determined that 42 students could be selected and examined, with the funds available, to estimate the parameters of the problem given in example 5.1. Using equal allocation method, he selected n_h = n/L = 42/3 = 14 students from each stratum by using WOR simple random sampling. The information so obtained from the selected students is given in the following table :

Table 5.3 Time (in hours) devoted to study in university library during a week

Stratum I		Stratum II		Stratum III	
0	10	7	14	15	24
2	0	8	6	17	14
1	7	11	4	9	8
3	8	5	6	18	20
5	3	9	12	24	11
6	8	10	6	22	21
8	4	12	13	23	16

Estimate the parameters of example 5.1 from the above data.

Solution

Using the data given in table 5.3 above, we prepare the following table :

Table 5.4 Values of various statistics calculated from data given in table 5.3

Stratum I	Stratum II	Stratum III
$n_1 = 14$	$n_2 = 14$	$n_3 = 14$
$N_1 = 1300$	$N_2 = 450$	$N_3 = 250$
$W_1 = .650$	$W_2 = .225$	$W_3 = .125$
$\bar{y}_1 = 4.643$	$\bar{y}_2 = 8.786$	$\bar{y}_3 = 17.286$
$s_1^2 = 10.707$	$s_2^2 = 10.484$	$s_3^2 = 29.132$

From expression (5.1) and table 5.4

$$\bar{y}_{st} = \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3)$$

$$= \frac{1}{2000}[1300(4.643) + 450(8.786) + 250(17.286)]$$

$$= 7.156$$

is the estimate of the weekly average time, in hours, devoted to study by a student in PAU library. Also from (5.3), the estimate of variance is

$$v(\bar{y}_{st}) = \frac{W_1^2(N_1 - n_1)s_1^2}{N_1 n_1} + \frac{W_2^2(N_2 - n_2)s_2^2}{N_2 n_2} + \frac{W_3^2(N_3 - n_3)s_3^2}{N_3 n_3}$$

$$= \frac{(.650)^2(1300 - 14)(10.707)}{(1300)(14)} + \frac{(.225)^2(450 - 14)(10.484)}{(450)(14)}$$

$$+ \frac{(.125)^2 (250-14)(29.132)}{(250)(14)}$$

$$= .3196 + .0367 + .0307$$

$$= .3870$$

Using (2.8), we obtain the lower and upper limits of the confidence interval as

$$\bar{y}_{st} \pm 2\sqrt{v(\bar{y}_{st})}$$

$$= 7.156 \pm 2\sqrt{.3870}$$

$$= 5.912, 8.400$$

To summarize, the estimate of the average time per week devoted to study by a student in PAU library is 7.156 hours. We are confident, with probability approximately equal to .95, that the actual average library study time per week for the PAU students will lie between 5.912 and 8.400 hours. ■

It may be noted that in case of equal allocation, no population characteristic is taken into consideration for determining the sample sizes for different strata. One needs to know only the number of strata to be constructed for finding the values of n_h, h = 1,2,..., L. However, it seems only reasonable that the importance of characteristics like strata sizes, variability, and per unit cost of observing study variable, which may change from stratum to stratum, be recognized and given due weight while determining the sample sizes $\{n_h\}$. The methods of sample allocation that we shall discuss now, are based on these considerations.

5.4.2 *Proportional Allocation*

This allocation was first proposed by Bowley (1926). When no other information except N_h, h = 1,2,...,L , is available, the size of strata is taken into account, and the number of units are drawn in proportion to the size of strata. This means $n_h \propto N_h$, implying that $n_h = (n/N)N_h$, h = 1,2,...,L. On substituting in (5.6) the value of n_h thus obtained, one gets the total sample size that can be selected and observed with the available money.

Sample size for h-th stratum in case of proportional allocation :

$$n_h = \frac{n}{N} N_h \qquad (5.9)$$

Total sample size for fixed total cost :

$$n = \frac{C - c_0}{\sum_{h=1}^{L} W_h c_h} \qquad (5.10)$$

Because of its simplicity, this procedure of allocation is often resorted to in practice. The allocation is likely to be nearly optimum for a fixed sample size, when the strata variances are almost same. On using the allocation $n_h = (n/N)N_h$, h = 1,2,...,L, in (5.2) to (5.5), one gets corresponding expressions for variance and estimator of variance for the estimator of population mean under proportional allocation.

Example 5.4
The third student of the group of four, was independently assigned the estimation problem of example 5.1, and was asked to use proportional allocation method . Using the budget and cost information of example 5.2, determine the total number of students that he could afford to select. Also, allocate the sample units to different strata.

Solution
In this case, we have $N_1 = 1300$, $N_2 = 450$, $N_3 = 250$, N = 2000, L = 3, C = \$150, $c_0 = \$24$, and $c_1 = c_2 = c_3 = \$3$. The total sample size that could be possible with the given information, is obtained by using (5.10). As $W_h = N_h/N$, the expression (5.10) can be written as

$$n = \frac{N(C - c_0)}{\sum_{h=1}^{L} N_h c_h}$$

$$= \frac{(2000)(150 - 24)}{(1300)(3) + (450)(3) + (250)(3)}$$

$$= 42$$

The total number of 42 students to be included in the sample are allocated to each of the 3 strata through (5.9). Thus,

$$n_1 = \left(\frac{n}{N}\right) N_1 = \left(\frac{42}{2000}\right)(1300) = 27.3 \approx 27$$

$$n_2 = \left(\frac{n}{N}\right) N_2 = \left(\frac{42}{2000}\right)(450) = 9.5 \approx 10$$

$$n_3 = \left(\frac{n}{N}\right) N_3 = \left(\frac{42}{2000}\right)(250) = 5.3 \approx 5$$

Therefore, 27, 10, and 5 students would be selected from strata I, II, and III respectively. ■

Example 5.5
In order to estimate the parameters of example 5.1, the total sample size that could be possible with the given budget has been obtained, in example 5.4, as 42 along with its proportional allocation to different strata. Accordingly, the student investigator selected 27 students from stratum I, 10 from stratum II, and 5 from stratum III. The information collected from the students in the sample is given in table 5.5.

Table 5.5 Time (in hours) devoted to study in library during a week

Stratum I					Stratum II		Stratum III
4	5	11	3	2	5	12	18
3	6	0	8	1	9	8	20
10	4	1	2	7	7	10	17
6	9	10	4		16	17	23
8	3	5	6		11	7	10
1	12	4	5				

Estimate the parameters of example 5.1.

Solution

For computations, we prepare table 5.6.

Table 5.6 Calculated values of various statistics for data given in table 5.5

Stratum I		Stratum II		Stratum III	
n_1 =	27	n_2 =	10	n_3 =	5
N_1 =	1300	N_2 =	450	N_3 =	250
W_1 =	.650	W_2 =	.225	W_3 =	.125
$\bar{y}_1$ =	5.185	$\bar{y}_2$ =	10.200	$\bar{y}_3$ =	17.600
s_1^2 =	10.851	s_2^2 =	15.289	s_3^2 =	23.300

Using (5.1), and various values from table 5.6, we find

$$\bar{y}_{st} = \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3)$$

$$= \frac{1}{2000}[1300\,(5.185) + 450\,(10.200) + 250\,(17.600)]$$

$$= 7.865$$

The estimate of variance $V(\bar{y}_{st})$ is computed by using (5.3) as

$$v(\bar{y}_{st}) = \frac{W_1^2(N_1 - n_1)\,s_1^2}{N_1 n_1} + \frac{W_2^2\,(N_2 - n_2)\,s_2^2}{N_2 n_2} + \frac{W_3^2(N_3 - n_3)\,s_3^2}{N_3 n_3}$$

$$= \frac{(.650)^2\,(1300 - 27)\,(10.851)}{(1300)\,(27)} + \frac{(.225)^2\,(450 - 10)\,(15.289)}{(450)\,(10)}$$

$$+ \frac{(.125)^2\,(250 - 5)\,(23.300)}{(250)\,(5)}$$

$$= .1663 + .0757 + .0714$$

$$= .3134$$

Utilizing (2.8), we work out the confidence interval from

$$\bar{y}_{st} \pm 2\sqrt{v(\bar{y}_{st})}$$
$$= 7.865 \pm 2\sqrt{.3134}$$
$$= 6.745, 8.985$$

Thus, on the average, a PAU student devotes 7.865 hours per week to study in the library. So far as the actual population average is concerned, we are reasonably confident that it will fall in the closed interval [6.745, 8.985] hours. ■

5.4.3 *Optimum/Neyman Allocation*

Very often, a survey statistician has to work within a fixed budget. In such a situation, he is expected to minimize the variance of the estimator subject to the cost constraint. In certain other cases, from the point of view of the results of the survey, it might be possible to state an acceptable value of the variance. The problem then is to minimize the cost, subject to the constraint that the variance of the estimator does not exceed the stated value. First we consider the former situation.

Case I. In this case, the total cost of the survey is fixed. The aim is to find the sample allocation $\{n_h\}$ such that the variance of the estimator is minimum. The allocation $\{n_h\}$, which minimizes the variance in (5.2) for a given cost C in (5.6), is called *optimum allocation*. The sample allocation, so obtained, is given in (5.11). The overall sample size for the optimum allocation can, however, be found by adding the n_h values obtained under this allocation.

Fixed total cost - minimum variance allocation :

$$n_h = \frac{(C - c_0)\, W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}} \tag{5.11}$$

Total sample size :

$$n = \sum_{h=1}^{L} n_h \tag{5.12}$$

where n_h has been given in (5.11).

We thus see that the stratum sample size will be proportional to the stratum size and stratum standard deviation, but inversely proportional to the square root of the cost per sampling unit in each stratum. It means, large strata with greater variability and low per unit observation cost will lead to larger samples in relation to those from other strata. In case the sampling is WR, the sample allocation $\{n_h\}$ that minimizes the variance in (5.4) is obtained by replacing S_h by σ_h in (5.11).

Example 5.6

A car manufacturing company has sold 2000 cars to the public through licensed dealers. The company is now interested in finding out the average distance travelled per week by a car manufactured by the company. This information is likely to be helpful in fixing the warranty period for certain parts of the car. The addresses and telephone numbers, if installed, of all the buyers along with their occupations are available at the head office of the company. Since the distance travelled by a car is likely to vary with the profession of the buyer, the investigator divides the population into 3 groups - the businessmen (stratum I), employees (stratum II), and others (stratum III) which includes farmers, etc. Out of 2000 buyers, 825 are businessmen, 700 employees, and 475 others. The average per unit cost for collecting information is expected to be \$ 4 for businessmen, \$ 5.5 for employees, and \$ 6.5 for persons from other category. The total budget at hand is \$ 1550 which includes the overhead cost of \$1000. On using optimum allocation formula given in (5.11), the investigator arrived at allocation of sample size $n_1 = 53$ buyers to stratum I, $n_2 = 34$ buyers to stratum II, and $n_3 = 23$ buyers to stratum III (procedure of determining these allocations is explained in the solution). The observations on the study variable obtained from these three WOR simple random samples are given in table 5.7.

Table 5.7 Average distance (in km) per week covered by cars included in the sample

Stratum I				Stratum II			Stratum III	
656	301	575	666	470	281	685	712	236
400	870	525	715	351	410	492	679	824
526	813	310	691	625	240	206	665	385
774	861	650	480	388	636	579	319	650
780	722	470	680	566	422	358	840	585
812	705	460	841	421	517	385	421	496
805	831	483	825	398	451		666	704
525	748	310	488	881	380		848	569
401	446	489	330	434	326		410	614
806	856	576	580	405	595		549	
828	387	615	811	693	401		602	
746	399	704		615	612		253	
560	635	774		375	564		777	
475	560	533		469	343		411	

The information on strata mean squares, from a similar survey carried out in the past for another car model, is given for strata I, II, and III respectively as $S_1^2 = 30505$, $S_2^2 = 24008$, and $S_3^2 = 29215$.

Solution

Here we have

$$C = \$\ 1550,\ c_o = \$\ 1000,\ c_1 = \$\ 4,\ c_2 = \$\ 5.5,\ c_3 = \$\ 6.5,$$
$$N_1 = 825,\ N_2 = 700,\ N_3 = 475,\ N = 2000,\ W_1 = .4125,\ W_2 = .3500,$$
$$W_3 = .2375,\ S_1^2 = 30505,\ S_2^2 = 24008,\ \text{and}\ S_3^2 = 29215.$$

The sample size allocation to different strata from (5.11) will be

$$n_h = \frac{(C - c_0)\, W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}},\quad h = 1,2,\ldots, L$$

Now, we first compute

$$\sum_{h=1}^{L} W_h S_h \sqrt{c_h} = (.4125)(\sqrt{30505})(\sqrt{4}) + (.3500)(\sqrt{24008})(\sqrt{5.5}) + (.2375)(\sqrt{29215})(\sqrt{6.5})$$
$$= 374.77$$

Then, the sample size allocation for the three strata will be

$$n_1 = \frac{(1550 - 1000)(.4125)\sqrt{30505}}{(374.77)\sqrt{4}} = 52.87 \approx 53$$

$$n_2 = \frac{(1550 - 1000)(.3500)\sqrt{24008}}{(374.77)\sqrt{5.5}} = 33.94 \approx 34$$

$$n_3 = \frac{(1550 - 1000)(.2375)\sqrt{29215}}{(374.77)\sqrt{6.5}} = 23.37 \approx 23$$

The information collected from WOR simple random samples of 53, 34, and 23 respondents selected from strata I, II, and III is given in table 5.7 above. The sample means and sample mean squares for the 3 strata are then calculated. These are given below :

$$\bar{y}_1 = 619.038 \qquad \bar{y}_2 = 469.824 \qquad \bar{y}_3 = 574.565$$
$$s_1^2 = 28531.190 \qquad s_2^2 = 20696.634 \qquad s_3^2 = 32871.256$$

The estimate of mean is now computed from (5.1). Thus,

$$\bar{y}_{st} = (W_1\bar{y}_1 + W_2\bar{y}_2 + W_3\bar{y}_3)$$
$$= \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3)$$
$$= \frac{1}{2000}[(825)(619.038) + (700)(469.824) + (475)(574.565)]$$
$$= 556.251$$

Hence, the estimate of the average distance per weak covered by a car, manufactured by the company, is 556.251 km.

For calculating the estimate of variance we use (5.3), and get

$$v(\bar{y}_{st}) = \frac{W_1^2(N_1 - n_1)\, s_1^2}{N_1 n_1} + \frac{W_2^2\,(N_2 - n_2)\, s_2^2}{N_2 n_2} + \frac{W_3^2(N_3 - n_3)\, s_3^2}{N_3 n_3}$$

$$= \frac{(.4125)^2\ (825 - 53)\ (28531.190)}{(825)\ (53)} + \frac{(.3500)^2\ (700 - 34)\ (20696.634)}{(700)\ (34)}$$

$$+ \frac{(.2375)^2\ (475 - 23)\ (32871.256)}{(475)\ (23)}$$

$$= 85.7147 + 70.9468 + 76.7115$$

$$= 233.373$$

Further, we have

$$\bar{y}_{st} \pm 2\sqrt{v(\bar{y}_{st})}$$

$$= 556.251 \pm 2\sqrt{233.373}$$

$$= 525.698,\ 586.804$$

Thus, the confidence interval for the population value of the average distance covered per week by a car is obtained as [525.698, 586.804] km. ■

If the cost per unit is same for all the strata, that is, $c_h = c'$ for each h, optimum allocation is known as *Neyman allocation*, after Neyman (1934). For this case, the cost function (5.6) takes more simpler form

$$C = c_0 + c'n \tag{5.13}$$

Then, the strata sample sizes for Neyman allocation are given as :

Minimum variance – Neyman allocation :

$$n_h = n\,\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} = n\,\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \tag{5.14}$$

Total sample size :

$$n = \frac{C - c_o}{c'} \tag{5.15}$$

Example 5.7

The fourth student, in the group of four, was asked to undertake the estimation of parameters considered in examples 5.1 to 5.5 by using Neyman allocation. Again, the cost for contacting the students and gathering the required information is $ 3 per student. The total budget at disposal was $150 including the overhead cost of $24. Using Neyman allocation, the samples of sizes 25, 10, and 7 students were selected from strata I, II, and III respectively (procedure of determining these sample sizes is explained in the solution). The data collected from these three WOR simple random samples, selected from three strata, are presented in table 5.8. Using this information, estimate the average time per week devoted to study in library by a student. Also, set up confidence interval for the population mean. The information on strata mean squares is to be used from example 5.5.

Solution

We are given that

$C = \$150$, $c_o = \$24$, $c' = \$\ 3$, $N_1 = 1300$, $N_2 = 450$, $N_3 = 250$, $S_1^2 = 10.851$, $S_2^2 = 15.289$, and $S_3^2 = 23.300$.

Taking cost into account, the total sample size from (5.15) will be

$$n = \frac{C - c_o}{c'} = \frac{150 - 24}{3} = 42$$

Now,

$$\sum_{h=1}^{3} N_h S_h = N_1 S_1 + N_2 S_2 + N_3 S_3$$

$$= (1300)\left(\sqrt{10.851}\right) + (450)\left(\sqrt{15.289}\right) + (250)\left(\sqrt{23.300}\right)$$

$$= 7248.615$$

The sample sizes for different strata are then determined using (5.14), where

$$n_h = n \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}, \quad h = 1, 2, \ldots, L$$

Thus,

$$n_1 = (42) \frac{(1300)\left(\sqrt{10.851}\right)}{7248.615} = 24.81 \approx 25$$

$$n_2 = (42) \frac{(450)\left(\sqrt{15.289}\right)}{7248.615} = 10.20 \approx 10$$

$$n_3 = (42) \frac{(250)\left(\sqrt{23.300}\right)}{7248.615} = 6.99 \approx 7$$

The observations recorded from these selected students are given below in table 5.8.

Table 5.8 Time (in hours) devoted to study in library by selected students during a week

Stratum I					Stratum II		Stratum III	
9	6	8	6	10	9	16	24	25
1	7	3	9	5	14	6	18	22
3	2	5	2	3	13	8	11	
5	4	4	5	4	8	12	19	
4	6	4	1	7	12	11	16	

For convenience, we compute sample estimates for mean and mean square error for each stratum. These are given along with other required information in table 5.9.

Table 5.9 Necessary computations for strata I, II, and III

Stratum I	Stratum II	Stratum III
$n_1 = 25$	$n_2 = 10$	$n_3 = 7$
$N_1 = 1300$	$N_2 = 450$	$N_3 = 250$
$W_1 = .650$	$W_2 = .225$	$W_3 = .125$
$\bar{y}_1 = 4.920$	$\bar{y}_2 = 10.900$	$\bar{y}_3 = 19.286$
$s_1^2 = 5.993$	$s_2^2 = 9.656$	$s_3^2 = 23.892$

By using figures from table 5.9 in (5.1), we work out the estimate of weekly average study time in the library as

$$\bar{y}_{st} = \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3)$$

$$= \frac{1}{2000}[1300\,(4.920) + 450\,(10.900) + 250\,(19.286)]$$

$$= 8.061$$

Next we compute estimate of variance $V(\bar{y}_{st})$ from (5.3) as

$$v(\bar{y}_{st}) = \frac{W_1^2(N_1 - n_1)\,s_1^2}{N_1 n_1} + \frac{W_2^2\,(N_2 - n_2)\,s_2^2}{N_2 n_2} + \frac{W_3^2(N_3 - n_3)\,s_3^2}{N_3 n_3}$$

$$= \frac{(.650)^2\,(1300 - 25)\,(5.993)}{(1300)\,(25)} + \frac{(.225)^2\,(450 - 10)\,(9.656)}{(450)\,(10)}$$

$$+ \frac{(.125)^2\,(250 - 7)\,(23.892)}{(250)\,(7)}$$

$$= .0993 + .0478 + .0518$$

$$= .1989$$

The required confidence interval for population mean is then obtained from

$$\bar{y}_{st} \pm 2\sqrt{v(\bar{y}_{st})}$$
$$= 8.061 \pm 2\sqrt{.1989}$$
$$= 7.169, 8.953$$

One can claim that there is approximately 95% chance that the population value of the average weekly study time in the university library will be covered by the closed interval [7.169, 8.953] hours. ■

Case II. Here, we fix the precision of the estimator at a specified level and minimize the total cost of survey. The desired level of precision can be specified in two ways. It could be done either by fixing the value of variance $V(\bar{y}_{st})$ at V_o, or by specifying the bound B on the error for the estimator $\bar{y}_{st}$. These two modes of precision specification are, however, related to each other through the equation

$$2\sqrt{V(\bar{y}_{st})} = B$$

implying

$$V(\bar{y}_{st}) = V_o = \frac{B^2}{4}$$

Above mentioned relation can, therefore, be used to convert one mode of specification to the other. Thus for any given value of B, we can find the corresponding value of V_o. So, if it is required to ensure a specified value V_o of the variance of the estimator, then on using the cost function (5.6), the sample size for the h-th stratum, h=1,2,..,L, is found as the one given in (5.16). The total sample size is again the sum of n_h values obtained under this allocation.

Fixed variance - minimum cost allocation :

$$n_h = \frac{\left(W_h S_h / \sqrt{c_h}\right) \sum_{h=1}^{L} W_h S_h \sqrt{c_h}}{V_o + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2} \qquad (5.16)$$

where V_o is the value at which the variance of the estimator $\bar{y}_{st}$ is fixed.

Total sample size :

$$n = \sum_{h=1}^{L} n_h \qquad (5.17)$$

where n_h has been obtained in (5.16).

For the cost function in (5.13), the above allocation reduces to minimum cost-Neyman allocation.

Minimum cost - Neyman allocation :

$$n_h = \frac{(W_h S_h) \sum_{h=1}^{L} W_h S_h}{V_o + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2} \qquad (5.18)$$

Total sample size :

$$n = \sum_{h=1}^{L} n_h \qquad (5.19)$$

where n_h has been obtained in (5.18).

In order to obtain sample allocation and total sample size for estimation of mean/total for a given variance V_o in case of WR sampling, take the factor $(\Sigma W_h S^2_h)/N$, h=1,2,...,L, as zero and replace S_h^2 by σ_h^2 in (5.16) and (5.18).

In both the cases (fixed cost or fixed variance) of optimum/ Neyman allocation, it is necessary to have approximate values for S_h^2 (or σ_h^2) for each stratum to be able to use the allocation formulas. These values will not be normally available in practice. Approximate values of these strata mean squares can be had from some similar studies conducted in the past, or from the ranges of the observations within each stratum, if available. The range of a set of data is known to be approximately four times its standard deviation. Alternatively, a pilot survey, based on a small sample, could also be conducted to obtain sample estimates s_h^2, of S_h^2 in case of WOR sampling and of σ_h^2 if the sampling is to be with replacement. These estimated values of S_h^2 (or σ_h^2) can then be used to determine sample sizes for different strata.

Example 5.8

If the car manufacturer wishes to estimate the parameters of example 5.6 with a predetermined variance V_o = 170, what will be the total sample size and allocation of sample to different strata, so that, the total cost is minimum for the same per unit costs as in example 5.6 ? The estimates of strata mean squares (s_1^2 = 28531.190, s_2^2 = 20696.634, and s_3^2 = 32871.256) obtained in example 5.6 are to be used as known strata mean squares for the purpose of present allocation.

Solution

From the statement of the example, we have V_o = 170. The other information from example 5.6 is

c_1 = \$4, c_2 = \$5.5, c_3 = \$6.5, N_1 = 825, N_2 = 700,

N_3 = 475, W_1 = .4125, W_2 = .3500, and W_3 = .2375.

The estimated values of S_h^2 in example 5.6 are now taken as good approximations of S_h^2, h = 1, 2, 3. Therefore,

$S_1^2 = 28531.190$, $S_2^2 = 20696.634$, and $S_3^2 = 32871.256$.

For a fixed variance and minimum cost, the sample allocation is given by (5.16) as

$$n_h = \frac{\left(W_h S_h / \sqrt{c_h}\right) \sum_{h=1}^{L} W_h S_h \sqrt{c_h}}{V_o + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2}$$

First, we calculate two terms that are to be used later on. These terms are

$$\begin{aligned}\sum_{h=1}^{L} W_h S_h \sqrt{c_h} &= (.4125)(\sqrt{28531.190})(\sqrt{4}) + (.3500)(\sqrt{20696.634})(\sqrt{5.5}) \\ &\quad + (.2375)(\sqrt{32871.256})(\sqrt{6.5}) \\ &= 367.220\end{aligned}$$

$$\begin{aligned}\sum_{h=1}^{L} W_h S_h^2 &= (.4125)(28531.190) + (.3500)(20696.634) \\ &\quad + (.2375)(32871.256) \\ &= 26819.861\end{aligned}$$

It follows that

$$\begin{aligned}V_o + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2 &= 170 + \frac{26819.861}{2000} \\ &= 183.410\end{aligned}$$

Then, the sample allocation is seen to be

$$n_1 = \frac{(.4125)(\sqrt{28531.190})(367.220)}{(\sqrt{4})(183.410)} = 69.8 \approx 70$$

$$n_2 = \frac{(.3500)(\sqrt{20696.634})(367.220)}{(\sqrt{5.5})(183.410)} = 43$$

$$n_3 = \frac{(.2375)(\sqrt{32871.256})(367.220)}{(\sqrt{6.5})(183.410)} = 33.8 \approx 34$$

The total sample size required will, therefore, be

$$\begin{aligned}n &= n_1 + n_2 + n_3 \\ &= 70 + 43 + 34 \\ &= 147 \blacksquare\end{aligned}$$

After discussing the various popular methods of sample allocation to different strata, we now attempt to answer the question whether a particular stratification and sample allocation combination will at all be advantageous in relation to the unstratified simple random sampling ?

5.5 RELATIVE EFFICIENCY OF STRATIFIED ESTIMATOR

For examining the usefulness of stratification, we need the sampling variances of the estimators of population mean/total for stratified and unstratified population. The percent relative efficiency of the estimator $\bar{y}_{st}$, with respect to the usual unstratified estimator $\bar{y}$, is then given by

$$RE = \frac{V(\bar{y})}{V(\bar{y}_{st})} (100)$$

where $V(\bar{y})$ and $V(\bar{y}_{st})$, for WOR sampling, are defined in (3.9) and (5.2) respectively. We illustrate below, the various steps involved in the calculation of the above said actual percent relative efficiency for three commonly used sample allocation methods. A relative efficiency figure of well over 100 indicates that the stratification of the population would be effective in reducing the estimation error. For this purpose, we consider a hypothetical situation where the study variable values are known for all population units.

Example 5.9
All the 80 farms in a population are stratified by farm size. The expenditure on the insecticides used during the last year by each farmer is presented in table 5.10 below :

Table 5.10 Expenditure (in '00 rupees) on insecticides used

Large farmers		Medium farmers				Small farmers		
75	76	55	40	51	28	35	31	26
65	79	45	38	55	47	28	38	32
86	62	35	33	41	61	36	42	18
57	92	30	43	48	35	40	33	16
45	50	42	53	54	31	25	29	
69	48	38	37	36	23	18	25	
48	77	40	52	44		28	35	
60	60	36	39	47		32	26	
55	64	48	46	39		13	30	
66	58	46	42	41		19	37	

Select a stratified sample of 24 farmers by using equal allocation, proportional allocation, and Neyman allocation. Compute the overall population mean $\bar{Y}$ and the population mean square S^2. Work out the relative efficiency of stratified sample mean $\bar{y}_{st}$, based on each of the above mentioned allocations, with respect to the simple random sample mean $\bar{y}$ for the same total sample size. Assume that the sampling is WOR.

Solution

It is given that $n = 24$, $N_1 = 20$, $N_2 = 36$, and $N_3 = 24$. Hence $W_1 = .25$, $W_2 = .45$, and $W_3 = .30$. First, we calculate overall population mean $\bar{Y}$ and population mean square S^2 based on all the 80 farms. Thus,

$$\bar{Y} = \frac{1}{80}(75+65+...+16)$$

$$= 43.7875$$

$$S^2 = \frac{1}{80-1}[(75)^2 + (65)^2 + ...+(16)^2 - (80)(43.7875)^2]$$

$$= 268.6758$$

From (3.9), the sampling variance of mean in case of usual SRS without replacement is given by

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

$$= \frac{80-24}{(80)(24)}(268.6758)$$

$$= 7.8364$$

Analogous to S^2, the stratum mean square S_h^2 is computed separately for each stratum. Hence, one gets

$S_1^2 = 169.5158$, $S_2^2 = 70.5611$, and $S_3^2 = 61.4493$.

We now obtain the value of $V(\bar{y}_{st})$ under three sample allocation methods.

Equal allocation. In this case, the number of units to be selected from each stratum will be $n_h = 24/3 = 8$. The sampling variance for stratified mean $\bar{y}_{st}$ from (5.2), will be

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) S_h^2$$

$$= (.25)^2 \left(\frac{20-8}{(20)(8)}\right)(169.5158) + (.45)^2 \left(\frac{36-8}{(36)(8)}\right)(70.5611)$$

$$+ (.30)^2 \left(\frac{24-8}{(24)(8)}\right)(61.4493)$$

$$= .7946 + 1.3892 + .4609$$

$$= 2.6447$$

The percent relative efficiency of equal allocation based mean $\bar{y}_{st}$, with respect to usual mean $\bar{y}$, is obtained as

$$RE = \frac{V(\bar{y})}{V(\bar{y}_{st})}(100)$$

$$= \frac{7.8364}{2.6447}(100)$$

$$= 296.3$$

Proportional allocation. The number of units to be selected from each stratum, using proportional allocation defined in (5.9), will be

$$n_1 = \left(\frac{n}{N}\right)N_1 = \left(\frac{24}{80}\right)(20) = 6$$

$$n_2 = \left(\frac{n}{N}\right)N_2 = \left(\frac{24}{80}\right)(36) = 10.8 \approx 11$$

$$n_3 = \left(\frac{n}{N}\right)N_3 = \left(\frac{24}{80}\right)(24) = 7.2 \approx 7$$

In place of sample size 8, used for equal allocation for calculating variance $V(\bar{y}_{st})$, here we use 6, 11, and 7 for strata I, II, and III respectively. Therefore,

$$V(\bar{y}_{st}) = (.25)^2\left(\frac{20-6}{(20)(6)}\right)(169.5158) + (.45)^2\left(\frac{36-11}{(36)(11)}\right)(70.5611)$$

$$+ (.30)^2\left(\frac{24-7}{(24)(7)}\right)(61.4493)$$

$$= 1.2361 + .9021 + .5596$$

$$= 2.6978$$

The percent relative efficiency of proportional allocation based stratified mean estimator $\bar{y}_{st}$, with respect to usual mean estimator $\bar{y}$, is given by

$$RE = \frac{7.8364}{2.6978}(100)$$

$$= 290.5$$

Neyman allocation. To arrive at the number of units to be selected from each stratum under Neyman allocation, we first work out $\Sigma N_h S_h$, h = 1, 2, 3. Thus,

$$\Sigma N_h S_h = (20)(\sqrt{169.5158}) + (36)(\sqrt{70.5611}) + (24)(\sqrt{61.4493})$$

$$= 260.4 + 302.4 + 188.1$$

$$= 750.9$$

Then from (5.14), for h = 1, 2 ,..., L, we have

$$n_1 = n\frac{N_1S_1}{\Sigma N_hS_h} = (24)\frac{260.4}{750.9} = 8.3 \approx 8$$

$$n_2 = n\frac{N_2S_2}{\Sigma N_hS_h} = (24)\frac{302.4}{750.9} = 9.7 \approx 10$$

$$n_3 = n\frac{N_3S_3}{\Sigma N_hS_h} = (24)\frac{188.1}{750.9} = 6$$

On using the above allocated sample sizes, the variance of mean estimator becomes

$$V(\bar{y}_{st}) = (.25)^2\left(\frac{20-8}{(20)(8)}\right)(169.5158) + (.45)^2\left(\frac{36-10}{(36)(10)}\right)(70.5611)$$

$$+ (.30)^2\left(\frac{24-6}{(24)(6)}\right)(61.4493)$$

$$= .7946 + 1.0320 + .6913$$

$$= 2.5179$$

The percent relative efficiency of Neyman allocation based mean estimator, with respect to usual simple random sampling estimator of mean, is obtained as

$$RE = \frac{7.8364}{2.5179}(100)$$

$$= 311.2 \blacksquare$$

It will be observed in the above example, that the calculated $\{n_h\}$ values often involve approximations due to the rounding off to the nearest integer. These approximations ultimately affect the sampling variance. The more exact values of the variance $V(\bar{y}_{st})$ can be obtained by using alternative forms of variance expressions. These forms of variances do not directly involve $\{n_h\}$ values. Hence no rounding off approximations are involved. These expressions can be obtained by putting respective $\{n_h\}$ values (in expression form) in (5.2). We shall here consider such variance expressions only for proportional and Neyman allocation methods. These expressions, given in (5.20) and (5.21), can respectively be obtained by substituting $\{n_h\}$ values from (5.9) and (5.14) in (5.2). Similar variance expression for the equal allocation method can be easily obtained by putting $n_h = n/L$, h = 1, 2,..., L, in (5.2). The reader must note that such type of alternative expressions can not be obtained for estimators of variance $V(\bar{y}_{st})$, since each s_h^2 has to be calculated from samples of sizes that are whole numbers.

Example 5.10

For the data of example 5.9, work out variances of stratified sample mean $\bar{y}_{st}$, under proportional and Neyman allocations, by using the expressions

$$V_p(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{L} W_h S_h^2 \tag{5.20}$$

$$V_n(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \tag{5.21}$$

respectively.

Solution

Using different values already computed in example 5.9, we calculate the following two terms :

$$\sum_{h=1}^{L} W_h S_h^2 = (.25)(169.5158) + (.45)(70.5611) + (.30)(61.4493)$$

$$= 92.5662$$

$$\sum_{h=1}^{L} W_h S_h = (.25)(\sqrt{169.5158}) + (.45)(\sqrt{70.5611}) + (.30)(\sqrt{61.4493})$$

$$= 9.3867$$

On substituting the different values in expressions of $V_p(\bar{y}_{st})$ and $V_n(\bar{y}_{st})$, one gets

$$V_p(\bar{y}_{st}) = \left(\frac{1}{24} - \frac{1}{80}\right)(92.5662)$$

$$= 2.6998$$

$$V_n(\bar{y}_{st}) = \frac{1}{24}(9.3867)^2 - \frac{1}{80}(92.5662)$$

$$= 2.5142 \blacksquare$$

For examples 5.9 and 5.10, it was assumed that the values of the study variable are available for all population units. In practice, however, the situation is different. The investigator has observations on the study variable for the stratified sample only. It is from this sample data that one has to estimate the variances $V(\bar{y}_{st})$ and $V(\bar{y})$. These estimated variances are then used to estimate the relative efficiency of the estimator $\bar{y}_{st}$ with respect to the estimator $\bar{y}$.

The estimator of $V(\bar{y}_{st})$ from a stratified simple random WOR sample is available in (5.3), while the estimator $v_{st}(\bar{y})$ of $V(\bar{y})$ from the stratified sample is obtained in (5.22) as

$$v_{st}(\bar{y}) = \frac{N-n}{Nn(N-1)}\left[\sum_{h=1}^{L} \frac{N_h}{n_h}\left(\sum_{i=1}^{n_h} y_{hi}^2\right) - N\{\bar{y}_{st}^2 - v(\bar{y}_{st})\}\right] \tag{5.22}$$

where $\bar{y}_{st}$ and $v(\bar{y}_{st})$ are defined in (5.1) and (5.3) respectively.

The estimated percent relative efficiency is then given by

$$RE = \frac{v_{st}(\bar{y})}{v(\bar{y}_{st})} (100)$$

Various steps involved in calculating the estimate of RE are explained in example 5.11.

Example 5.11

The sample data obtained by using proportional allocation are given in example 5.5. Estimate the relative efficiency of proportional allocation based stratified estimator $\bar{y}_{st}$, in relation to usual unstratified simple mean estimator $\bar{y}$, from the above referred stratified sample observations.

Solution

From example 5.5, we have $N_1 = 1300$, $N_2 = 450$, $N_3 = 250$, $N = 2000$, $n_1 = 27$, $n_2 = 10$, $n_3 = 5$, $\bar{y}_{st} = 7.865$, and $v(\bar{y}_{st}) = .3133$. Now from table 5.5, we compute the term $\sum_{i=1}^{n_h} y_{hi}^2$ for h = 1,2,3. Thus,

$$\sum_{i=1}^{27} y_{1i}^2 = 4^2 + 3^2 + \ldots + 7^2 = 1008$$

$$\sum_{i=1}^{10} y_{2i}^2 = 5^2 + 9^2 + \ldots + 7^2 = 1178$$

$$\sum_{i=1}^{5} y_{3i}^2 = 18^2 + 20^2 + \ldots + 10^2 = 1642$$

Using above computed values, we calculate

$$\sum_{h=1}^{L} \frac{N_h}{n_h} \left(\sum_{i=1}^{n_h} y_{hi}^2 \right) = \frac{1300}{27} (1008) + \frac{450}{10} (1178) + \frac{250}{5} (1642)$$

$$= 183643.33$$

Now on making substitutions in (5.22), we obtain the estimated variance of the usual SRS estimator $\bar{y}$ from the stratified sample. Thus,

$$v_{st}(\bar{y}) = \frac{2000 - 42}{(2000)(42)(2000-1)} [183643.33 - (2000)\{(7.865)^2 - .3133\}]$$

$$= .7061$$

The required percent relative efficiency is then estimated as

$$RE = \frac{v_{st}(\bar{y})}{v(\bar{y}_{st})} (100)$$

$$= \frac{.7061}{.3133}(100)$$

$$= 225.38 \blacksquare$$

5.6 ESTIMATION OF POPULATION PROPORTION

So far, we have dealt with the estimation of population mean $\bar{Y}$ and population total Y on the basis of with and without replacement stratified simple random samples. The results for the estimation of $\bar{Y}$ can easily be extended for the estimation of population proportion P. For this, we take the value of y_{hi} as 1 or 0 according as the unit belongs to the class of interest or not. In this case, $\bar{y}_h$, $\bar{Y}_h$, and $\bar{Y}$ reduce to h-th stratum sample proportion p_h, the h-th stratum proportion P_h, and the overall population proportion P respectively. One can also see that for this case $\sigma_h^2 = P_h Q_h$, where $Q_h = 1 - P_h$, h = 1, 2, ..., L. Thus we have (5.23).

Unbiased estimator of population proportion P :

$$p_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h p_h = \sum_{h=1}^{L} W_h p_h \qquad (5.23)$$

The expressions for variance $V(p_{st})$ and its estimator $v(p_{st})$, for stratified simple random sampling WOR, are given below :

Variance of estimator p_{st} :

$$V(p_{st}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h - 1}{N_h - 1}\right) \frac{P_h Q_h}{n_h} \qquad (5.24)$$

Estimator of variance $V(p_{st})$:

$$v(p_{st}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{p_h q_h}{n_h - 1} \qquad (5.25)$$

In case of WR sampling, the fpc = $[1- (n_h-1) / (N_h-1)]$ and the sampling fraction $f_h = n_h/N_h$ are respectively taken as 1 and 0. The expressions (5.24) and (5.25) then reduce to their counterparts in WR method.

So far as the allocation and determination of sample size for estimation of population proportion P is concerned, the discussion of section 5.4 still holds, except for the difference that σ_h^2 and S_h^2 will be replaced by $P_h Q_h$ and $N_h P_h Q_h/(N_h-1)$ respectively, whereas s_h^2 will be replaced by $n_h p_h q_h/(n_h-1)$.

Example 5.12

The management of a local newspaper is to decide whether it should continue with the publication of 'Children Column', which had been introduced on experimental basis. For this purpose, it is imperative to estimate the proportion of readers who would favor its continuance. The frame consists of readers who had stayed with the paper for the last six months. The addresses of these readers are available in the office of the newspaper. Since different attitudes are expected from the urban and rural readers, it is reasonable to stratify the population into urban readers and rural readers. In the population, there are 73000 urban readers and 30280 rural readers. The total budget at hand is \$3000 only. The overhead cost is \$820, and the per unit cost for urban and rural readers is expected to be \$2 and \$2.5 respectively. Using proportional allocation method, explained in solution, the investigator selected WOR simple random samples of 718 respondents from stratum I (urban readers) and 298 readers from stratum II (rural readers). The number of individuals who favor continuation of the column was 570 from stratum I and 143 from stratum II. Estimate the proportion of readers interested in the continuation of the said column. Also, build up confidence interval for the population proportion.

Solution

Here we have,

$N_1 = 73000$, $N_2 = 30280$, $N = 103280$, $C = \$3000$, $c_o = \$820$,
$c_1 = \$2$, $c_2 = \$2.5$, $W_1 = N_1/N = .7068$, and $W_2 = N_2/N = .2932$.

Total number of respondents who can be surveyed is given by (5.10) as

$$n = \frac{C - c_o}{\sum_{h=1}^{L} W_h c_h} = \frac{(3000 - 820)}{(.7068)(2) + (.2932)(2.5)} = 1015.6 \approx 1016$$

Using (5.9), the number of respondents to be selected from stratum I and stratum II are

$$n_1 = \left(\frac{n}{N}\right) N_1 = \left(\frac{1016}{103280}\right)(73000) = 718.1 \approx 718$$

$$n_2 = \left(\frac{n}{N}\right) N_2 = \left(\frac{1016}{103280}\right)(30280) = 297.9 \approx 298$$

The estimate of proportion of urban readers who are in favor of continuing the 'Children Column' is

$$p_1 = \frac{570}{718} = .7939$$

Similarly, the estimate of proportion of rural readers, who wish the continuance of the said column, is

$$p_2 = \frac{143}{298} = .4799$$

The estimate of the required overall population proportion is now worked out from (5.23) as

$$p_{st} = \frac{1}{N}(N_1p_1 + N_2p_2)$$

$$= \frac{1}{103280}[(73000)(.7939) + (30280)(.4799)]$$

$$= .7018$$

Next, we compute estimate of variance from (5.25). It yields

$$v(p_{st}) = W_1^2\left(\frac{N_1 - n_1}{N_1}\right)\frac{p_1q_1}{n_1 - 1} + W_2^2\left(\frac{N_2 - n_2}{N_2}\right)\frac{p_2q_2}{n_2 - 1}$$

$$= (.7068)^2\left(\frac{73000 - 718}{73000}\right)\left(\frac{.7939\ (1 - .7939)}{718 - 1}\right)$$

$$+ (.2932)^2\left(\frac{30280 - 298}{30280}\right)\left(\frac{.4799\ (1 - .4799)}{298 - 1}\right)$$

$$= .0001129 + .0000715$$

$$= .0001844$$

The confidence interval for population proportion is worked out by using (2.8) as

$$p_{st} \pm 2\sqrt{v(p_{st})}$$

$$= .7018 \pm 2\sqrt{.0001844}$$

$$= .6746, .7290$$

To summarize, the sample estimate of proportion indicates that about 70 percent of reader population is in favor of continuing the column in question. The confidence limits obtained above assure the management of the paper that the population proportion favoring the continuance of the column will lie in the closed interval [.6746, .7290] with probability approximately equal to .95. ■

Example 5.13

After 6 months of the survey considered in example 5.12, the management of that newspaper again wishes to estimate the proportion of readers who would favor continuance of 'Children Column'. The budget at disposal, overhead cost, and per unit cost for collecting information for urban and rural readers remain same as in the survey considered in example 5.12. The estimates of strata variances obtained in example 5.12 are to be used as known strata variances for the present survey for the purpose of sample allocation. Using all this information, allocate the sample to different strata following optimum allocation method.

Solution

From example 5.12, we have $N_1 = 73000$, $N_2 = 30280$, $N = 103280$, $C = \$3000$, $c_0 = \$820$, $c_1 = \$2$, $c_2 = \$2.5$, $W_1 = .7068$, and $W_2 = .2932$. Also,

$$A_1 = \frac{n_1 p_1 q_1}{n_1 - 1} = \frac{718\,(.7939)\,(1-.7939)}{718-1} = .1639$$

$$A_2 = \frac{n_2 p_2 q_2}{n_2 - 1} = \frac{298\,(.4799)\,(1-.4799)}{298-1} = .2504$$

The optimum allocation for quantitative data is given by (5.11). By using the estimate s_h in place of S_h, the expression is written as

$$n_h = \frac{(C - c_0) W_h s_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h s_h \sqrt{c_h}}$$

In case of attributes, corresponding to s_h^2 we have $\frac{n_h p_h q_h}{n_h - 1}$. The above allocation formula, therefore, reduces to

$$n_h = \frac{(C - c_0) W_h \sqrt{A_h} / \sqrt{c_h}}{\sum_{h=1}^{L} W_h \sqrt{A_h} \sqrt{c_h}}, \quad h = 1, 2$$

where A_1 and A_2 have been defined and calculated above. Now, we work out

$$\sum_{h=1}^{L} W_h \sqrt{A_h} \sqrt{c_h} = .7068\,(\sqrt{.1639})\,(\sqrt{2}) + .2932\,(\sqrt{.2504})\,(\sqrt{2.5})$$

$$= .6367$$

The substitution of different values in the above formula for n_h gives

$$n_1 = \frac{(3000 - 820)\,(.7068)\,(\sqrt{.1639})}{(\sqrt{2})\,(.6367)} = 692.8 \approx 693$$

$$n_2 = \frac{(3000 - 820)\,(.2932)\,(\sqrt{.2504})}{(\sqrt{2.5})\,(.6367)} = 317.7 \approx 318$$

Therefore, under optimum allocation that minimizes the variance, 693 respondents will be selected from urban readers and 318 from the rural stratum. ■

5.7 CONSTRUCTION OF STRATA

As pointed out earlier, the basic consideration involved in the formation of strata is that the strata should be internally homogeneous. For a single study variable y, the best characteristic for construction of strata is the distribution of study variable y itself. L strata could then be formed by cutting this distribution at (L-1) suitable points. This distribution of y is generally not available in practice, and in absence of this information, the next

best alternative is the frequency distribution of some other variable which is highly correlated with the study variable y. Construction of strata on such an auxiliary variable will not yield exactly optimum strata, but these will be approximately optimum. In what follows, the procedures of constructing strata for different allocation methods are discussed.

5.7.1 *Neyman and Equal Allocation*

Dalenius and Hodges (1957) gave cumulative square root rule to obtain approximately optimum strata for Neyman allocation by using the frequency distribution for the study variable y. As this distribution is not available in practice, the rule is used on the frequency distribution of a highly positively correlated auxiliary variable x (also called stratification variable) to obtain approximately optimum stratification on x. Cumulative square root rule, though proposed for the Neyman allocation method, is also found to yield approximately optimum strata for equal allocation method. This rule can, therefore, be used for the construction of strata for both Neyman and equal allocation methods. Steps involved in the construction of strata, through the above rule, are listed as under :

Steps involved in cumulative square root rule :

1. Obtain a frequency table for stratification variable x.
2. In the frequency table for x, obtain square roots of the frequencies for each of the K classes.
3. Obtain the cumulative totals of the square roots of frequencies for each of the K classes. Let T denote the cumulative total for the K-th class.
4. If L strata are to be constructed, then using linear interpolation method on the class intervals and the cumulative square root frequency column, obtain the value of $x = x_1$, which corresponds to the value T/L in the cumulative square root frequency column.
5. Repeat the process in step (4) to obtain $x = x_i$ corresponding to the value i.T/L, i = 2,3,...,L–1, in the cumulative square root frequency column.
6. The values $(x_1, x_2,..., x_{L-1})$ so obtained define L strata with boundaries $(< x_1)$, $(x_1$ to $x_2)$, $(x_2$ to $x_3)$,...,$(x_{L-2}$ to $x_{L-1})$, and $(\geq x_{L-1})$.

Example 5.14

It is desired to estimate average annual milk yield per cow for a *tharparkar* herd of 127 cows at a certain government cattle farm using stratified simple random sampling. Cows in the herd are to be grouped into three strata on the basis of first lactation length in days. Neyman method of sample allocation is to be used for selecting the overall sample of 25 cows from the three strata. Determine approximately optimum strata boundaries using the information on first lactation length given in table 5.11.

Table 5.11 First lactation length (in days) and other related computations

Lactation length	No. of cows (f)	$\sqrt{f}$	Cumulative $\sqrt{f}$
30 - 70	4	2.00	2.00
70 - 110	6	2.45	4.45
110-150	3	1.73	6.18
150-190	8	2.83	9.01
190-230	20	4.47	13.48
230-270	27	5.20	18.68
270-310	25	5.00	23.68
310-350	14	3.74	27.42
350-390	7	2.65	30.07
390-430	6	2.45	32.52
430-470	6	2.45	34.97
470-510	1	1.00	35.97

Solution

In this example, we are already given the frequency table for the stratification variable, first lactation length. As the next step, we find square roots of the frequencies (f) given in column (2) of the table 5.11. These square root values ($\sqrt{f}$) are presented in column (3). The cumulative totals of $\sqrt{f}$ are then obtained. These totals constitute column (4) of the table.

For this illustration, we have L=3, K=12, and T=35.97. For constructing three strata, we need to determine only two boundaries, x_1 and x_2 in days, using linear interpolation between the class intervals and the cumulative $\sqrt{f}$ values. As stated earlier, x_1 and x_2 are to correspond to T/3 = 35.97/3 = 11.99 and 2T/3 = 2(35.97)/3 = 23.98 in column (4). From table 5.11, we find that a value of 9.01 in column (4) corresponds to the value 190 in column (1), whereas a value of 13.48 in column (4) corresponds to the value 230 in column (1). Thus, an increase of 4.47 in cumulative $\sqrt{f}$ value takes place over the interval 190-230. First lactation length x_1 corresponding to the cumulative $\sqrt{f}$ value of 11.99, therefore, lies in this interval. Hence,

$$x_1 = 190 + \frac{(40)\,(11.99 - 9.01)}{4.47}$$

$$= 216.67$$

Similarly,

$$x_2 = 310 + \frac{(40)\,(23.98 - 23.68)}{3.74}$$

$$= 313.21$$

It shows that the cows with the first lactation length in the range [30, 216.67] will constitute the first stratum, whereas those having lactation lengths in the ranges [216.67, 313.21] and [313.21, 510] will form second and third strata respectively. ■

5.7.2 *Proportional Allocation*

Singh (1975) proposed a cumulative cube root rule to obtain approximately optimum strata boundaries for the proportional allocation method. The rule is to be applied on the frequency table for the stratification variable x in exactly the same way as the cumulative square root rule of Dalenius and Hodges (1957), except for the difference that in this method we shall operate with the cube roots of the class frequencies in place of their square roots.

Example 5.15

It is proposed to estimate total wool yield in a certain region of Rajasthan state in India, using stratified simple random sampling. An overall sample of 20 villages is to be selected employing proportional allocation method. The stationary sheep population data, for 141 villages of this region, is given in the frequency table below. Construct three approximately optimum strata taking stationary sheep population as the stratification variable.

Table 5.12 Stationary sheep population with other related computations

No. of sheep	No. of villages (f)	$\sqrt[3]{f}$	Cumulative $\sqrt[3]{f}$
0-100	46	3.583	3.583
100-200	36	3.302	6.885
200-300	23	2.844	9.729
300-400	11	2.224	11.953
400-500	6	1.817	13.770
500-600	4	1.587	15.357
600-700	4	1.587	16.944
700-800	1	1.000	17.944
800-900	4	1.587	19.531
900-1000	4	1.587	21.118
1000-1100	1	1.000	22.118
1100-1200	1	1.000	23.118

Solution

For this case, the cube roots of frequencies in column (2) are given in column (3), and their cumulated values are presented in column (4) of table 5.12. Here we have L=3, K=12, and T= 23.118. The two strata boundaries, x_1 and x_2, needed to form three strata will now correspond to T/3 = 23.118/3 = 7.706 and 2T/3 = 2(23.118)/3 =15.412 respectively. Thus, on proceeding as in example 5.14, we get from columns (1) and (4)

$$x_1 = 200 + \frac{(100)(7.706 - 6.885)}{2.844} = 228.87$$

$$x_2 = 600 + \frac{(100)(15.412 - 15.357)}{1.587} = 603.47$$

Hence, the villages having stationary sheep population in the range [0, 228.87] constitute first stratum, and those having sheep population in the ranges [228.87, 603.47] and [603.47, 1200] form second and third strata respectively. ■

5.8 POSTSTRATIFICATION

In stratified sampling, it is presupposed that the strata sizes and the sampling frame for each stratum are available. However, the situations do exist where the latter is difficult to obtain. For instance, the details about classification of farmers' population by farm size (small, medium, large) can be had from the census records, but list of farmers falling in each of the three classes may not be available. Consequently, it is not possible to determine in advance as to which stratum a farmer belongs until he is observed for the farm size. This means, the units can be assigned to different strata only after the sample units are contacted and observed. This whole procedure is termed as *poststratification*. This technique is useful where published journals/ reports may provide clear indication of strata sizes, but due to nonavailability of strata frames it is difficult to sample the units from different strata. Below we give the estimator of population mean, its variance, and estimator of this variance when the sample units are stratified after they have been selected as a single WOR simple random sample from the entire unstratified population.

Estimator of population mean $\bar{Y}$:

$$\bar{y}_{ps} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{5.26}$$

Approximate variance of estimator $\bar{y}_{ps}$:

$$V(\bar{y}_{ps}) = \frac{N-n}{Nn} \sum_{h=1}^{L} W_h S_h^2 + \frac{1}{n^2} \sum_{h=1}^{L} (1-W_h) S_h^2 \tag{5.27}$$

Estimator of variance $V(\bar{y}_{ps})$:

$$v(\bar{y}_{ps}) = \sum_{h=1}^{L} \left(\frac{N_h - n_h}{N_h n_h} \right) W_h^2 s_h^2 \tag{5.28}$$

The first term of (5.27) is the value of $V(\bar{y}_{st})$ for proportional allocation. The second term is due to the fact that $\{n_h\}$ do not distribute themselves exactly proportionally because of poststratification. However, for sufficiently large n, this second term will be small in comparison to first. It means that for reasonably large n, the poststratification is almost as precise as stratified sampling with proportional allocation.

Example 5.16

The list of all the 800 farmers in a development block can be obtained but the information about their farm sizes is not available. Thus the farmers can not be classified as large, medium, or small, and the usual stratified sampling procedure is not applicable. However, from census records the proportion of large, medium, and small farmers in

the above said population is known to be .2, .5, and .3 respectively. The State Bank of India is interested in estimating the total amount of loan expected to be taken by the farmers of the development block during the next financial year. For this purpose, a WOR simple random sample of 40 farmers was selected and then classified according to their farm sizes. The information regarding the expected amount of loan to be taken by the selected farmers is given below :

Table 5.13 Expected amount of loan (in '000 rupees) to be taken by the sample farmers

Large farmers	Medium farmers			Small farmers	
44	30	29	35	18	16
50	42	30	33	22	20
60	28	29	36	17	14
38	20	34	27	26	23
52	30	25	31	21	28
43	31	27	24	19	24
				16	15
				10	16

Estimate the total amount of loan expected to be taken by all the farmers in the block. Also, place confidence limits on this total.

Solution

Here, we have $W_1 = .2$, $W_2 = .5$, $W_3 = .3$, and $N = 800$. Also, $n_1 = 6$, $n_2 = 18$, and $n_3 = 16$. Now proceeding as in examples 5.1 and 5.3, we compute sample means and sample mean squares for each stratum. These are given below along with values of N_h, W_h, and n_h. The N_h values for h=1, 2, 3, are computed from the equation $N_h = NW_h$.

Table 5.14 Certain sample and population values

Large farmers	Medium farmers	Small farmers
$n_1 = 6$	$n_2 = 18$	$n_3 = 16$
$W_1 = .2$	$W_2 = .5$	$W_3 = .3$
$N_1 = 160$	$N_2 = 400$	$N_3 = 240$
$\bar{y}_1 = 47.833$	$\bar{y}_2 = 30.056$	$\bar{y}_3 = 19.063$
$s_1^2 = 60.967$	$s_2^2 = 24.526$	$s_3^2 = 22.596$

From (5.26), the estimate of the total amount of loan, expected to be taken by all the farmers of the block, will be

$$\hat{Y} = N\bar{y}_{ps} = \sum_{h=1}^{L} N_h \bar{y}_h$$
$$= 160\,(47.833) + 400\,(30.056) + 240\,(19.063)$$
$$= 24250.80$$

We now compute estimate for the variance of $\hat{Y}$. For this, we use (5.28). Thus,

$$v(\hat{Y}_{ps}) = N^2\, v(\bar{y}_{ps}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h} s_h^2$$
$$= \frac{160\,(160-6)}{6}\,(60.967) + \frac{400\,(400-18)}{18}\,(24.526)$$
$$+ \frac{240\,(240-16)}{16}\,(22.596)$$
$$= 250371.14 + 208198.48 + 75922.56$$
$$= 534492.18$$

The required confidence limits are worked out following (2.8). These are given by

$$\hat{Y}_{ps} \pm 2\sqrt{v(\hat{Y}_{ps})}$$
$$= 24250.80 \pm 2\sqrt{534492.18}$$
$$= 22788.62,\ 25712.98$$

The investigator could, therefore, be reasonably confident that the total amount of loan, expected to be taken by all the 800 farmers of the block, is in the range of 22788.62 to 25712.98 thousand rupees. ■

While dealing with poststratification, we have assumed that strata sizes are known exactly. But in certain situations, it may not be the case. In such a situation, one is either to use guess values of $\{N_h\}$ from some past survey or census record, or one can estimate them by drawing a large preliminary sample and then use a subsample to obtain information on the study variable. In both the cases, the values of $\{N_h\}$ used may differ from the actual $\{N_h\}$. These inaccuracies affect the precision of the required estimate. Detailed discussion of the problem is given in Sukhatme *et al.* (1984).

5.9 SOME FURTHER REMARKS

5.1 Suppose a sample of n_h units is selected from N_h units of the h-th stratum with PPS with replacement using x as the size variable. Let Y_{hi} and $P_{hi} = (X_{hi}/X_h)$ respectively, denote the study variable value and the probability of selection for the i-th unit of the h-th stratum. Also, let y_{hi} and p_{hi} be the corresponding sample values. Then we have the following results :

Unbiased estimator of population total Y:

$$\hat{Y}_{pst} = \sum_{h=1}^{L} \hat{Y}_h = \sum_{h=1}^{L} \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{y_{hi}}{p_{hi}} \tag{5.29}$$

Variance of estimator $\hat{Y}_{pst}$:

$$V(\hat{Y}_{pst}) = \sum_{h=1}^{L} V(\hat{Y}_h) = \sum_{h=1}^{L} \frac{1}{n_h} \sum_{i=1}^{N_h} \left(\frac{Y_{hi}}{P_{hi}} - Y_h \right)^2 P_{hi} \tag{5.30}.$$

Estimator of variance $V(\hat{Y}_{pst})$:

$$v(\hat{Y}_{pst}) = \sum_{h=1}^{L} v(\hat{Y}_h) = \sum_{h=1}^{L} \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} \left(\frac{y_{hi}^2}{p_{hi}^2} - n_h \hat{Y}_h^2 \right) \tag{5.31}$$

where $\hat{Y}_h$ is defined in (5.29).

5.2 In certain situations, the population is highly heterogeneous and several effective criteria are available for stratification. In such cases, it may be desirable to carry stratification to the extent that only one unit is selected from each stratum. In this event, it is not possible to estimate $V(\bar{y}_{st})$ and $V(\hat{Y}_{st})$. However, approximate estimate of variance is possible by using method of *collapsed strata*. With L even, this method consists in grouping the adjoining strata in pairs to form collapsed strata, and then estimating the sampling variance as if two units had been sampled from each collapsed stratum. Hansen, Hurwitz, and Madow (1953) have given a more general procedure by grouping the strata into g groups.

5.3 Goodman and Kish (1950) proposed that it is possible to enhance the control of error further. The method they suggested is termed *controlled selection* procedure. It increases the probability of selection of preferred samples, and, consequently, reduces the probability of selection of nonpreferred samples without altering the probabilities of selection to the units in a stratified sampling design.

5.4 Sometimes, the investigator is interested in estimating population characteristics for several variables. If all the variables of interest are closely related to a single auxiliary variable, say x, and information for all the population units on x is available, then stratification and allocation is done as discussed in this chapter. If all the study variables are not related to a single auxiliary variable but are related to more than one auxiliary variable, the procedure of stratification and allocation is little modified and termed *multiple stratification* or *deep stratification*. In this procedure, the population units are first stratified using the most important auxiliary variable. The strata thus formed are called *primary strata*. Then each primary strata is further stratified using another auxiliary variable. For details, the reader may refer to Sukhatme *et al.* (1984).

5.5 Certain other methods of determining approximately optimum strata boundaries on the study variables have been proposed by Aoyama (1954), Dalenius and Gurney (1951), Dalenius and Hodges (1959), and Mahalanobis (1952). These

methods are, however, not preferred to cumulative square root rule. Methods of approximately optimum stratification on an auxiliary variable for certain other allocation procedures have been given by Singh (1971), Singh and Parkash (1975), and Mehta *et al.* (1995).

LET US DO

5.1 Explain, why should one use stratified simple random sampling ? What are the various points one should keep in mind while stratifying a population ?

5.2 Discuss various problems that are to be resolved before one could start selecting a stratified simple random sample.

5.3 How does stratification increase efficiency of the estimator of mean/total and proportion ?

5.4 Do you think it appropriate to use stratified sampling to estimate

a. per head travelling allowance (TA) of teaching and nonteaching staff of a university,
b. average calories taken per day by boys and girls,
c. proportion of nonexistent voters in a voter list, and
d. mean weight of adult men and women in a city block.

5.5 What do you understand by sample allocation? Describe merits and demerits of various sample allocation methods.

5.6 A stratified sample of size 80 is to be drawn from a population of 6400 units, divided into 3 strata of sizes 2400, 3200, and 800 units. If the allocation is to be equal, or proportional, how many units should be selected from individual stratum in each case?

5.7 An insurance company's records show that out of the total of 500 claims, 280 are major claims (from Rs 1000 to Rs 2500) and 220 are minor (below Rs 1000). A WOR simple random sample of 10 claims was drawn from each category (stratum), and claim amounts were recorded as :

Stratum I : 1200, 1600, 1800, 1400, 1980, 2110, 2440, 1660, 1790, 1910
Stratum II : 720, 880, 760, 660, 790, 840, 550, 960, 640, 800

Estimate the total amount of all the 500 claims, and construct the confidence interval for it.

5.8 The adult population in a colony consists of 400 Sikhs, 260 Muslims, 200 Hindus, and 140 Christians. An investigator selected 40 Sikhs, 26 Muslims, 18 Hindus, and 16 Christians so as to draw a total sample of 100 adults. Do you think the allocation is proportional ?

5.9 During 1990-91 session, a student doing M.Sc. (Statistics) was given a project to estimate average time taken by the university employees to get ready for office in the morning. The population was grouped into 3 strata. The first stratum consisted of women. The males were divided into 2 strata – teachers and the other staff. A WOR random sample of 400 employees was drawn using proportional

allocation. The information on time (in hours) taken by selected respondents to be ready for office was collected. Below are given the sample average and sample mean square for each stratum along with the values of N_h and n_h.

Strata	N_h	n_h	$\bar{y}_h$	s^2_h
Women	1250	125	2.0	.1587
Teachers	2390	239	1.6	.2667
Others	360	36	1.2	.0811
Total	4000	400		

Estimate the average time taken to get ready for office, and place the required confidence limits on it.

5.10 Discuss the method of sample allocation to different strata when (a) total cost of survey is fixed and the aim is to minimize variance, (b) precision of the estimator $\bar{y}_{st}$ is fixed and the objective is to minimize survey cost.

5.11 Take the estimates of strata mean squares obtained in exercise 5.9 as known. Using this information along with strata sizes, determine Neyman allocation when the total budget at disposal is Rs 1000. Assume that the overhead cost, and the cost of eliciting and processing information per respondent, are Rs 100 and Rs 3 respectively.

5.12 In March 1992, another student of M.Sc., majoring in statistics, was assigned the job to estimate the parameters of exercise 5.9 using Neyman allocation. For this, $\{s^2_h\}$ values in exercise 5.9 were treated as actual strata mean squares. One thousand rupees were allotted to him for the completion of this job. The overhead cost was expected to be Rs 200, and the cost of collecting and processing the information per respondent was known as Rs 2. This way, the number of units to be selected from each stratum came out to be $n_1 = 109$, $n_2 = 269$, and $n_3 = 22$. The allocated number of individuals were drawn from each stratum by using SRS without replacement, and the requisite information on time (in hours) spent by each sample individual to get ready for office work was obtained. In table below are presented the sample average and sample mean square for each stratum, along with the values of N_h and n_h.

Stratum	N_h	n_h	$\bar{y}_h$	s^2_h
Women	1250	109	2.3	.1308
Teachers	2390	269	1.6	.2111
Others	360	22	1.4	.0905

Estimate the parameter of exercise 5.9, and construct confidence interval for it.

5.13 Below are given 50 population values divided into three strata :

Stratum I : 78, 87, 71, 88, 76, 99, 98, 86, 75, 92

Stratum II : 40, 46, 42, 58, 60, 55, 49, 54, 61, 48, 44, 52,
54, 46, 58, 57, 50, 48, 44, 50, 47, 50, 41, 43

Stratum III : 30, 28, 28, 24, 26, 21, 34, 28, 36, 27, 26, 32,
24, 25, 29, 23

Compute the overall population mean $\bar{Y}$ and the population mean square S^2. If one is to select WOR random sample of size 10, determine the appropriate stratum sample sizes using proportional allocation and Neyman allocation. Work out the relative efficiency of the stratified sample mean $\bar{y}_{st}$ for proportional and Neyman allocations, with respect to the usual SRS based mean estimator $\bar{y}$, for the same total sample size of 10.

5.14 The list of all the 50,000 adults in a town was available. In order to estimate proportion of literate adults (educated up to at least 8th grade), the population was stratified into 3 strata with respect to age. A WOR random sample of size 500 persons was drawn using proportional allocation. The sample size allocation to each stratum and the number of literate persons recorded in sample of size n_h, h = 1, 2, 3, are given below :

Age group (years)	Persons	n_h	Literate persons
20-40	25600	256	243
40-60	18100	181	144
60 and over	6300	63	31

Compute the estimate of proportion of literate persons in the town, and construct confidence interval for it.

5.15 The Electricity Department has 4 offices in a town. There are complaints of receiving faulty electricity bills by the consumers. The 4 offices were treated as strata and the bills issued during the last six months were to be scrutinized. An amount of Rs 1720 was allotted to carry out the survey. The overhead cost was expected to be Rs 100, and the cost of scrutinizing a bill, and processing the collected information, is Rs 3. In all, 540 units could be observed with these funds. For determining the sample size for each stratum, proportional allocation was used. The results obtained from the survey are reported as follows :

Office	Total bills issued	n_h	Faulty bills
1	7000	140	13
2	4280	86	7
3	9560	191	9
4	6160	123	10
Total	27000	540	

Estimate total number of faulty bills issued by all the 4 offices, and place confidence limits on it.

5.16 Describe how will you obtain approximately optimum strata boundaries for Neyman allocation ?

5.17 The objective of a survey is to estimate the total wheat production in a certain region having 220 farmers. The area (in hectares) under wheat in respect of each farmer is available from revenue records. Its distribution is given below :

Area	Farmers	Area	Farmers
0-2	6	10-12	41
2-4	13	12-14	30
4-6	14	14-16	16
6-8	36	16-18	11
8-10	50	18-20	3

Neyman allocation method is to be employed for selecting the overall sample of 30 farmers after dividing the population into 3 strata. Determine the approximately optimum strata boundaries using the cumulative square root method.

5.18 Describe cumulative cube root rule for constructing approximately optimum strata. For which allocation method, this rule is appropriate ?

5.19 An overall sample of 30 farmers is to be selected from 3 strata using proportional allocation for the estimation of parameter of exercise 5.17. Determine approximately optimum strata boundaries by using Singh's (1975) cumulative cube root method on the frequency table of exercise 5.17.

5.20 What do you understand by poststratification? Identify 4 situations where this technique could be useful. Give expressions for the estimator of total, its variance, and estimator of variance.

5.21 The objective of a study was to estimate mean fibre length of three newly developed strains of cotton (*Gossypium arboreum* L.). Incidentally, the seeds of these strains got mixed at the time of packing and transporting from the headquarters. Since no spare seed was available, the plants were raised using the

mixed seed. However, the strains could be distinguished if the plant characteristics were examined carefully by using laboratory equipment. On maturity, the total number of plants were counted as 1200. The proportion of plants of each variety was guessed from the number of each type of seed known before mixing. These were $W_1 = .60$, $W_2 = .25$, and $W_3 = .15$. An overall WOR random sample of 60 plants was selected. The selected plants were observed critically and assigned to the appropriate strain category. The fibre length recorded (in mm) for the sample plants is given below :

Strain I				Strain II		Strain III
18.5	17.0	16.9	18.3	19.6	20.4	21.6
17.4	18.3	17.4	17.9	19.4	20.6	20.7
19.3	18.1	16.8	17.6	20.3	19.3	21.4
18.0	19.6	17.0	17.1	19.8	18.4	21.4
17.6	17.5	18.5	18.0	18.6	19.1	20.5
17.6	17.4	18.8	19.4	18.9	17.9	22.3
18.2	18.4	18.1	19.2	19.5	20.5	21.7
17.5	19.1	17.1	18.3	18.4	18.3	22.2
17.0	17.3	16.8	17.3			

Estimate the mean fibre length, and place confidence limits on this average.

CHAPTER 6

Systematic Sampling

6.1 LINEAR SYSTEMATIC SAMPLING

In the preceding chapters, we have considered methods of sampling in which successive units are selected at random. In this chapter, an alternative sampling procedure is considered. The scheme, besides ensuring for each unit equal probability of inclusion in the sample, selects the whole sample with just one random number.

Definition 6.1 The method in which only the first unit is selected at random, the rest being automatically selected according to a predetermined pattern, is known as *systematic sampling.*

Several kinds of systematic sampling procedures are available in literature. These methods are appropriate for different situations. However, in this chapter we shall discuss only the commonly used sample selection methods, and also point out their advantages and disadvantages. One such method is known as *linear systematic* (LS) *sampling.*

Suppose we want to select a systematic sample of size n from a population consisting of N units. The method of LS sampling is employed when N is a multiple of n, that is, N=nk where k is an integer. For explaining the procedure, let us assume that the nk serial numbers of the population units in the frame are rearranged in k columns as follows :

1	2	3	...	r	...	k
k+1	k+2	k+3	...	k+r	...	2k
2k+1	2k+2	2k+3	...	2k+r	...	3k
.	.	.	.	.		
.	.	.	.	.		
.	.	.	.	.		
(n-1)k+1	(n-1)k+2	(n-1)k+3	...	(n-1)k+r	...	nk

Then, for selecting a systematic sample of n units, we select a random number r such that $1 \leq r \leq k$. The number r is called *random start*, and k is termed as the *sampling interval.* Starting with r, every k-th unit is included in the sample. This way, the population units with serial numbers r, r+k, ..., r+(n-1)k will constitute the sample. For example, let N = 100, n = 5 then k = 100/5 = 20. Suppose the random number chosen from 1 to 20 is 16. With 16 as random start, the units bearing serial numbers 16, 36, 56, 76, and 96 will be selected in the sample.

In this case, the systematic sampling amounts to grouping the N units into k samples of exactly n units each in a systematic manner, and selecting one of these samples with probability 1/k. From above, it is clear that each of the N units occurs once and only once in one of the k samples. It thus ensures equal probability of inclusion in the sample for every unit in the population.

The systematic sampling has the nice feature of operational convenience. It lies in the fact that the selection of the first unit determines the whole sample. This operation is easier to understand, and can be speedily executed in relation to simple random sampling. The sampling procedure is particularly suited for situations, where the sample is to be selected by the field staff themselves. Because of simplicity in the execution of systematic sampling, it would be easy to train persons in using it. Thus, it would be desirable to employ this procedure, whenever the sampling work has to be carried out by a large number of persons stationed in different areas.

Systematic samples are well spread over the population, and there is no risk that any large contiguous part of the population will be left unrepresented. This scheme provides more efficient estimates of population mean/total in comparison to simple random sampling for populations with *linear trend*, where study variable values in the population tend to be linearly related with the serial numbers of units in the frame . It is also efficient for populations with autocorrelation.

Systematic sampling should, however, be used with considerable care in case of periodicity in the population where the values of the study variable for the units listed in the frame, tend to increase and then decrease and increase again in a cyclic manner with a definite period. For *periodic populations*, if the sampling interval is an odd multiple of half the period of the cycle, systematic sampling provides estimator of mean with considerably small variance. On the other hand, if the sampling interval is an integral multiple of the period of the cycle, systematic sampling is no better than usual simple random sampling. The periodic variation is likely to occur when the population consists of groups of equal or approximately equal number of units, and the units within each group are arranged according to some definite sequence. For instance, in a population census, the households are the sampling units, and the individuals in the family are arranged according to a set pattern, such as, head of the family first, then his wife and their children in order of their age. In such a case, systematic sampling with an interval equal to the group size, or its multiple, may lead to inefficient estimates since the units selected in a sample will tend to be more or less similar in respect of the characteristic under study. Other situations, where the populations may exhibit periodicity, could be the flow of road traffic past a point over 24 hours of the day, or the store sales over seven days of the week. When estimating an average over a time period, a systematic sample daily at 6 p.m. would obviously be injudicious. Instead, the investigator should see that the time and week days are equally represented for such estimation situations. The point to stress is that one should carefully examine the possibility of existing periodicity in the population. If it exists, it can rather be helpful in reducing the variation in sample estimates. A serious disadvantage of this scheme lies in its use with populations having unforeseen periodicity, which may substantially contribute to the bias in the estimate of mean/total. Stephan *et al.* (1940) and Lahiri (1954) have discussed, in detail, the pitfalls involved in using systematic sampling in case of populations having cyclic

variation.

Another serious disadvantage of the sampling scheme is that the variance of the estimator can not be estimated unbiasedly from a single sample.

This sampling procedure has been found to be very useful in forest surveys for estimating the volume of timber. Systematic samples of boats returning from sea, are used for estimating the total catch of fish. The sampling scheme has also been used in milk yield surveys for estimating the lactation yield.

We now consider an example to illustrate the use of linear systematic sampling for selection of samples.

Example 6.1

An insurance company's claims, in dollars, for one day are 400, 600, 570, 960, 780, 800, 460, 650, 440, 530, 470, 810, 625, 510, and 700. List all possible systematic samples of size 3, that can be drawn from this set of claims using linear systematic sampling. Also, obtain corresponding sample means.

Solution

Here the population size N=15, and the size of the sample to be selected is n = 3. The sampling interval k will thus be 15/3 = 5. The random number r to be selected from 1 to k can, therefore, take any value in the closed interval [1, 5]. Each random start from 1 to 5 will yield corresponding systematic sample. In all, there will be k=5 possible samples. These are given below in table 6.1 along with their means.

Table 6.1 Possible systematic samples and their means

Random start (r)	Serial No. of sample units	y-values for sample units	Sample mean
1	(1, 6, 11)	400, 800, 470	556.67
2	(2, 7, 12)	600, 460, 810	623.33
3	(3, 8, 13)	570, 650, 625	615.00
4	(4, 9, 14)	960, 440, 510	636.67
5	(5, 10, 15)	780, 530, 700	670.00

■

In practice, it may often happen that $N \neq nk$. In this case, k is taken as an integer nearest to N/n. Proceeding as above, the scheme gives rise to samples of variable size. For example, in another case, we might have N=14 and n=5. Then k is to be taken as 3. The three possible samples for $1 \leq r \leq 3$ will consist of units with serial numbers (1, 4, 7, 10, 13), (2, 5, 8, 11, 14), and (3, 6, 9, 12). Thus, two samples have five units whereas the third has only four units. That means, the actual sample size may be different from the required one. In such situations, sample mean also does not remain unbiased for the population mean. These disadvantages can be overcome by using a sampling procedure, that is known as *circular systematic* (CS) *sampling.*

6.2 CIRCULAR SYSTEMATIC SAMPLING

This scheme can be used in both the cases, where N=nk or N≠nk. The method regards the N units as arranged round a circle, and consists in choosing a random start from 1 to N instead of from 1 to k, where k is the integral value nearest to N/n. The unit corresponding to this random start is the first unit included in the sample. Thereafter, every k-th unit, from those assumed arranged round the circle, is selected until a sample of n units is chosen. More concisely, if r is a random start, $1 \leq r \leq N$, then the units corresponding to the serial numbers

$\{r+jk\}$, if $r+jk \leq N$

and

$\{r+jk-N\}$, if $r+jk > N$,

$j = 0, 1, 2, ..., (n-1)$, will be selected in the sample. Theoretically, there is no problem in choosing any other smaller value of k, but it will only restrict the spread of the sample over a segment of the population. To illustrate, let N=14, n=5, and k be taken as 3. If random start r, $1 \leq r \leq 14$, is 7, then the units with serial numbers 7, 10, 13, 2, and 5 are included in the sample. Diagrammatically, this selection can be represented as below :

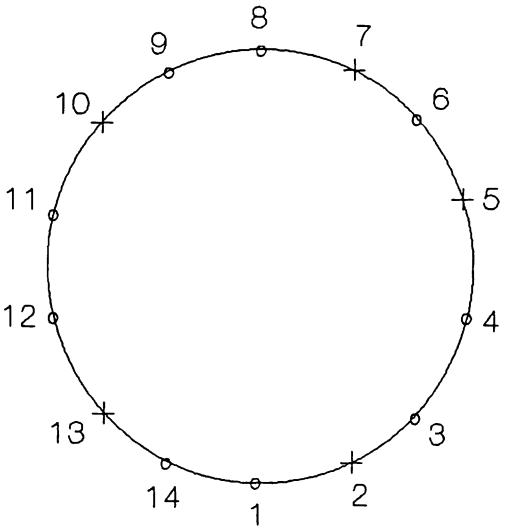

Fig. 6.1 Representation of CS sampling scheme

The CS sampling retains the two principal advantages: (1) it provides constant sample size, and (2) sample mean remains unbiased estimator of population mean. As mentioned earlier, it is not possible to obtain unbiased estimate of the sampling variance of the estimator from a single systematic sample. It remains a serious drawback of the circular systematic sampling procedure also.

Example 6.2
Using data of example 6.1, list all possible samples of size 4 along with their means, using circular systematic sampling.

Solution
We have N=15, n=4 and, therefore, k=4. In CS sampling, the random number r is selected from 1 to N. It will result in 15 random starts. Since corresponding to each random start there is one systematic sample, in all, one will obtain 15 possible systematic samples. These are listed in table 6.2 along with their means.

Table 6.2 Possible CS samples and their means

Random start (r)	Serial No. of sample units	y-values for sample units	Sample mean
1	(1, 5, 9, 13)	400, 780, 440, 625	561.25
2	(2, 6, 10, 14)	600, 800, 530, 510	610.00
3	(3, 7, 11, 15)	570, 460, 470, 700	550.00
4	(4, 8, 12, 1)	960, 650, 810, 400	705.00
5	(5, 9, 13, 2)	780, 440, 625, 600	611.25
6	(6, 10, 14, 3)	800, 530, 510, 570	602.50
7	(7, 11, 15, 4)	460, 470, 700, 960	647.50
8	(8, 12, 1, 5)	650, 810, 400, 780	660.00
9	(9, 13, 2, 6)	440, 625, 600, 800	616.25
10	(10, 14, 3, 7)	530, 510, 570, 460	517.50
11	(11, 15, 4, 8)	470, 700, 960, 650	695.00
12	(12, 1, 5, 9)	810, 400, 780, 440	607.50
13	(13, 2, 6, 10)	625, 600, 800, 530	638.75
14	(14, 3, 7, 11)	510, 570, 460, 470	502.50
15	(15, 4, 8, 12)	700, 960, 650, 810	780.00

■

6.3 ESTIMATING MEAN/ TOTAL

Before discussing the problem of estimation, let us assume that for the situation where N is not a multiple of n, the investigator uses CS sampling only. However, in case of N=nk, one may use either CS sampling or LS sampling. Under these assumptions, the sample mean is always unbiased for population mean. As mentioned earlier, an unbiased estimator of the variance of the sample mean is not available from a systematic sample with one random start, because a systematic sample could be regarded as a random sample of just one cluster (of units), and for estimating the variance one must have at least two such clusters in the sample. However, some biased estimators of variance are possible on the basis of a systematic sample. We consider one in (6.4), which takes into account successive differences of the sample values. However, if the units in the population are arranged at random then systematic sampling is equivalent to SRS without replacement.

In this case, the expression for variance estimator is same as in (3.10). For the sake of completeness, it is also given in (6.5).

Estimator of population mean :

$$\bar{y}_{sy} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad (6.1)$$

Variance of the estimator $\bar{y}_{sy}$:

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum_{r=1}^{k} (\bar{y}_{sy} - \bar{Y})_r^2 \qquad \text{(for LS sampling)} \qquad (6.2)$$

$$= \frac{1}{N} \sum_{r=1}^{N} (\bar{y}_{sy} - \bar{Y})_r^2 \qquad \text{(for CS sampling)} \qquad (6.3)$$

where $(\bar{y}_{sy}-\bar{Y})_r$ is the difference between the systematic sample mean corresponding to random start r and the population mean $\bar{Y}$.

Estimator of variance $V(\bar{y}_{sy})$:

$$v(\bar{y}_{sy}) = \frac{N-n}{2Nn(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2 \qquad (6.4)$$

$$v(\bar{y}_{sy}) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} (y_i - \bar{y}_{sy})^2 \qquad \text{(for random population)} \qquad (6.5)$$

As usual, the estimate of population total Y will be $\hat{Y}_{sy} = N\,\bar{y}_{sy}$. The variance and its estimator will be given by $V(\hat{Y}_{sy}) = N^2V(\bar{y}_{sy})$ and $v(\hat{Y}_{sy}) = N^2v(\bar{y}_{sy})$.

Example 6.3 (for N=nk)

About 70 years back, *Dalbergia sissoo* trees were planted in a single row on both sides of a road. The total number of trees are 3600. The Department of Public Works of a state is interested in estimating the total timber volume. A 1-in-100 systematic sample is selected. The data on estimated timber volume for the sampled trees (procedure of selection is given in solution) are presented in table 6.3. Estimate the total timber volume, and also construct the confidence interval for it.

Table 6.3 Timber volume (in cubic meters) for 36 selected trees

Serial No. of tree	Timber volume	Serial No. of tree	Timber volume	Serial No. of tree	Timber volume
28	1.72	1228	2.17	2428	1.89
128	1.29	1328	1.63	2528	1.63
228	1.08	1428	1.91	2628	2.23
328	2.29	1528	1.66	2728	2.40
428	2.01	1628	1.56	2828	2.51
528	1.77	1728	2.26	2928	2.57

Table 6.3 continued ...

Serial No. of tree	Timber volume	Serial No. of tree	Timber volume	Serial No. of tree	Timber volume
628	1.63	1828	2.49	3028	1.26
728	1.20	1928	2.26	3128	1.46
828	2.03	2028	2.31	3228	1.00
928	1.17	2128	1.60	3328	1.94
1028	2.47	2228	1.64	3428	1.80
1128	1.86	2328	1.43	3528	1.60

Solution

In this case, population size N=3600 and sampling interval k=100. We use linear systematic sampling for the selection of trees. Let the random number r selected from 1 to k(=100) be 28. The trees bearing serial numbers 28, 128, 228, 328, ..., 3528 will, therefore, be selected in the sample. The timber volume observed for the sample trees is given in table 6.3.

Estimate of total timber volume from (6.1) is

$$\hat{Y}_{sy} = N\bar{y}_{sy} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

$$= \frac{3600}{36}(1.72+1.29+...+1.60)$$

$$= \frac{3600}{36}(65.73)$$

$$= 6573$$

We now work out the estimate of variance $V(\hat{Y}_{sy})$ from (6.4) as

$$v(\hat{Y}_{sy}) = N^2 v(\bar{y}_{sy}) = \frac{N(N-n)}{2n(n-1)}\sum_{i=1}^{n-1}(y_{i+1}-y_i)^2$$

$$= \frac{N(N-n)}{2n(n-1)}[(y_2-y_1)^2 + (y_3-y_2)^2 + ... + (y_{36}-y_{35})^2]$$

$$= \frac{3600(3600-36)}{2(36)(35)}[(1.29-1.72)^2+(1.08-1.29)^2+ ...+(1.60-1.80)^2]$$

$$= \frac{3600(3600-36)}{2(36)(35)}(10.8056)$$

$$= 55015.94$$

Using the estimate for total timber volume and the estimate of its variance, we now calculate the confidence interval for population total from (2.8). It is given by

$$N\bar{y}_{sy} \pm 2N\sqrt{v(\bar{y}_{sy})}$$

$$= \hat{Y}_{sy} \pm 2\sqrt{v(\hat{Y}_{sy})}$$

$$= 6573 \pm 2\sqrt{55015.94}$$

$$= 6103.89, 7042.11$$

To summarize, the estimate of total timber volume obtained from the selected sample is 6573 cubic meters. It can be said with probability approximately equal to .95, that the actual total timber volume that can be had from all the 3600 trees, would be in the range of 6103.89 to 7042.11 cubic meters. ■

Example 6.4 (for N≠nk)

On a particular day, 162 boats had gone to sea from the coast for fishing. It was desired to estimate the total catch of fish at the end of the day. As it was not possible to weigh the catch for all the 162 boats, it was decided to weigh fish for only 15 boats selected using circular systematic sampling. Discuss the selection procedure, and obtain the estimate of total catch of fish using data on the 15 sample boats given in table 6.4.

Table 6.4 Catch of fish (in quintals) for 15 selected boats

Serial No. of boat	Catch of fish	Serial No. of boat	Catch of fish	Serial No. of boat	Catch of fish
73	5.614	128	9.225	21	8.460
84	8.202	139	6.640	32	10.850
95	6.115	150	7.350	43	6.970
106	9.765	161	5.843	54	5.524
117	8.550	10	6.875	65	7.847

Solution

In this case, we have N=162 and n=15. Since N/n=162/15=10.8 is not a whole number, the value of sampling interval k is taken as 11, an integer nearest to 10.8, and circular systematic sampling is used for selection of boats. If the selected random number r, $1 \le r \le 162$, is 73, then the boats bearing serial numbers 73, 84,..., 65 will be included in the sample. The serial numbers of selected boats, along with the corresponding catch of fish, are presented in table 6.4. We now proceed to estimate the total catch of fish using (6.1). This estimate is

$$\hat{Y}_{sy} = N\bar{y}_{sy} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

$$= \frac{162}{15}(5.614 + 8.202 + \ldots + 7.847)$$

$$= \frac{(162)(113.83)}{15}$$

$$= 1229.364$$

The estimate of variance $V(\hat{Y}_{sy})$ is then computed by using the expression (6.4). Thus,

$$v(\hat{Y}_{sy}) = N^2 v(\bar{y}_{sy}) = \frac{N(N-n)}{2n(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2$$

$$= \frac{162(162-15)}{2(15)(14)} [(8.202 - 5.614)^2 + (6.115 - 8.202)^2 + ... + (7.847 - 5.524)^2]$$

$$= \frac{(162)(162-15)(67.596)}{2(15)(14)}$$

$$= 3832.693$$

The confidence interval, for the total catch of fish for 162 boats, can then be calculated from

$$\hat{Y}_{sy} \pm 2\sqrt{v(\hat{Y}_{sy})}$$

$$= 1229.364 \pm 2\sqrt{3832.693}$$

$$= 1105.547, 1353.181$$

Thus, the estimate of total catch of fish obtained from a single sample is 1229.364 quintals. The confidence limits, obtained above, indicate that the total catch from all the 162 boats is likely to fall in the interval [1105.547, 1353.181] quintals. ■

The variance estimator given in (6.4) is biased and, therefore, should be used with care as inferences based on these estimates may sometimes be misleading in practice. Various approaches have been suggested to obtain unbiased variance estimators. Using a mixture of systematic and simple random sampling, Zinger (1963, 1964) suggested an unbiased estimator $\bar{y}_z$ of population mean but his proposed unbiased estimator of variance $V(\bar{y}_z)$ could not be proved nonnegative. Subsequently, Rana and Singh (1989) proposed another unbiased estimator of population mean, and also gave an unbiased estimator of variance of this estimator which was proved to be nonnegative. Discussion of these methods is, however, beyond the scope of this book. An alternative approach which provides unbiased estimator of variance, is through the use of interpenetrating subsampling. This approach we discuss in the following section.

6.4 ESTIMATING MEAN/TOTAL THROUGH INTERPENETRATING SUBSAMPLES

A method of estimating the variance $V(\bar{y}_{sy})$ unbiasedly, consists in selecting the sample of required size n in the form of two or more (say m) systematic subsamples of same size with independent random starts. Let us assume that these m *interpenetrating subsamples,* each of size n/m, are to be selected from the population of N units. Also, let N/n = k. Then, for selecting the required m samples using linear systematic sampling, we select m random starts, either using simple random sampling WR, or using simple random sampling WOR, from 1 to mk (assumed to be integer). The subsample corresponding to a particular random start will thus include population units with serial numbers at intervals of mk. In case mk is not an integer, we can use circular systematic

sampling for selecting the subsamples. Let $\bar{y}_1, \bar{y}_2, ..., \bar{y}_m$ be the estimators of the population mean based on m such subsamples, each of size n/m.

When m random starts are selected through simple random sampling WR, we have the following results :

Unbiased estimator of population mean $\bar{Y}$:

$$\bar{y}_{sy} = \frac{1}{m} \sum_{i=1}^{m} \bar{y}_i \tag{6.6}$$

Unbiased estimator of variance $V(\bar{y}_{sy})$:

$$v(\bar{y}_{sy}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} (\bar{y}_i - \bar{y}_{sy})^2 \tag{6.7}$$

If the sampling interval is small, selection of m random starts with replacement may lead to repetition of samples. In such cases, it is, therefore, desirable to select m interpenetrating systematic subsamples of n/m units each, with random starts selected from 1 to mk using without replacement sampling. This results in the selection of m of the mk possible samples with SRS without replacement. The formulas corresponding to this procedure are listed below :

Unbiased estimator of population mean $\bar{Y}$:

$\bar{y}_{sy}$ is same as in (6.6).

Variance of estimator $\bar{y}_{sy}$:

$$V(\bar{y}_{sy}) = \frac{k-1}{km(km-1)} \sum_{i=1}^{mk} (\bar{y}_i - \bar{Y})^2 \tag{6.8}$$

where $\bar{y}_i$ is the i-th subsample mean, i=1,2,...,mk.

Unbiased variance estimator :

$$v(\bar{y}_{sy}) = \frac{k-1}{km(m-1)} \sum_{i=1}^{m} (\bar{y}_i - \bar{y}_{sy})^2 \tag{6.9}$$

where $\bar{y}_1, \bar{y}_2, ..., \bar{y}_m$ are estimators based on m systematic samples with random starts selected with SRS without replacement.

Example 6.5

A dairy research institute is interested in estimating the total milk yield of a buffalo in connection with a breeding program. The milk yield of first five days was not recorded, being the colostrum period. The total lactation period was taken as 300 days. It was decided to select 3 systematic subsamples each of size 10 days, so as to arrive at a total

sample size of 30 days. The method of selecting the subsamples is discussed in the solution. Milk yield recorded for the selected days is given in table 6.5.

Table 6.5 Milk yield (in liters) for 30 selected days

Subsample I		Subsample II		Subsample III	
Selected days	Milk yield	Selected days	Milk yield	Selected days	Milk yield
1	8.10	12	9.30	26	11.15
31	12.00	42	13.50	56	14.70
61	15.20	72	14.40	86	14.60
91	14.00	102	14.35	116	12.80
121	11.25	132	9.80	146	10.65
151	10.10	162	10.00	176	10.60
181	9.80	192	8.60	206	8.30
211	8.75	222	8.40	236	7.50
241	7.25	252	6.10	266	4.30
271	4.10	282	3.10	296	2.20
Mean	10.055		9.755		9.680

Solution

In this problem, we have N=300, n=30, and m=3. This gives k=300/30=10. In order to avoid the possibility of repetition of subsamples, we select WOR three random starts from 1 to mk=30. Suppose we get the random starts as 1, 12, and 26. Corresponding to these three random starts, the selected days and corresponding milk yields are recorded in table 6.5. Subsample means are also given at the end of the table.

Estimate of total milk yield (in liters) can be obtained by using (6.6) as

$$\hat{Y}_{sy} = N\bar{y}_{sy} = \frac{N}{m}\sum_{i=1}^{m}\bar{y}_i$$

$$= \frac{300}{3}(10.055 + 9.755 + 9.680)$$

$$= 300\,(9.830)$$

$$= 2949$$

In this case, the estimate of variance is given by (6.9). Therefore,

$$v(\hat{Y}_{sy}) = N^2 v(\bar{y}_{sy})$$

$$= \frac{N^2(k-1)}{km(m-1)}\sum_{i=1}^{m}(\bar{y}_i - \bar{y}_{sy})^2$$

$$= \frac{(300)^2(10-1)}{10(3)(3-1)}[(10.055-9.830)^2 + (9.755-9.830)^2 + (9.680-9.830)^2]$$

$$= 1063.125$$

Confidence limits for total milk yield are derived from

$$\hat{Y}_{sy} \pm 2\sqrt{v(\hat{Y}_{sy})}$$

$$= 2949 \pm 2\sqrt{1063.125}$$

$$= 2883.789,\ 3014.211$$

The confidence limits computed above indicate that the total milk yield, if all the 300 observations were recorded, would most probably fall in the closed interval [2883.789, 3014.211] liters. ■

6.5 SAMPLE SIZE DETERMINATION FOR ESTIMATING MEAN/TOTAL

Assuming the population in random order, let n_1 be the number of units selected, in preliminary sample, from the population of N units. Then from the sampled n_1 units, we compute

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_i - \bar{y}_{sy1})^2 \qquad (6.10)$$

where

$$\bar{y}_{sy1} = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i$$

It may be noted that a systematic sample from a population in random order is equivalent to a simple random sample drawn without replacement. Therefore, the formulas for finding required sample size, obtained in section 3.5, will be applicable in this case also. These formulas are reproduced below for readers' convenience.

Sample size for estimating mean/total with a permissible error B :

$$n = \frac{Ns_1^2}{ND + s_1^2} \qquad (6.11)$$

where

$$D = \frac{B^2}{4} \quad \text{(when estimating mean)}$$

$$D = \frac{B^2}{4N^2} \quad \text{(when estimating total)}$$

with s_1^2 defined in (6.10). If $n_1 \geq n$, the sample size n_1 is sufficient, otherwise, $(n-n_1)$ additional units need to be selected.

As stated above, the assumption of population being in random order amounts to reducing systematic sampling to simple random sampling. Because of this, the expression in (6.11) above is same as the expressions in (3.18) and (3.20). In absence of this assumption, (6.11) could give an extra large sample for populations with linear trend and

too small a sample for periodic populations. Also, when the preliminary systematic sample of n_1 units is augmented by another systematic sample of $(n-n_1)$ units, the composite sample does not strictly remain a systematic sample.

Example 6.6

Assuming the population in example 6.3 to be in random order, and treating the sample of 36 trees selected there as the preliminary sample, determine the sample size required to estimate total timber volume with a tolerable error of 400 cubic meters.

Solution

Here, we are given B=400 cubic meters. From example 6.3, we have N=3600, n_1=36, and

$$\bar{y}_{sy1} = \frac{1}{36}(1.72+1.29+\ldots+1.60)$$
$$= 1.826$$

so that,

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_i-\bar{y}_{sy1})^2$$
$$= \frac{1}{n_1-1}\left(\sum_{i=1}^{n_1} y_i^2 - n_1\bar{y}_{sy1}^2\right)$$
$$= \frac{1}{36-1}[(1.72)^2+(1.29)^2+\ldots+(1.60)^2-36(1.826)^2]$$
$$= .1919$$

Also,

$$ND = \frac{B^2}{4N}$$
$$= \frac{(400)^2}{4(3600)} = 11.1111$$

Then, from (6.11), we can determine the required sample size as

$$n = \frac{Ns_1^2}{ND+s_1^2}$$
$$= \frac{(3600)(.1919)}{11.1111+(.1919)}$$
$$= 61.12$$
$$\approx 61$$

Since the sample size required to estimate the total timber volume with a permissible error of 400 cubic meters is 61, the investigator will, therefore, need to select 61-36=25 more trees to get the estimate with specified magnitude of tolerable error. ■

6.6 ESTIMATION OF PROPORTION

Sometimes, the investigator is interested in estimating the proportion P of population possessing a particular attribute, from a systematic sample. For instance, he/she may wish to estimate the proportion of voters who are satisfied with the functioning of Municipal Committee, an elected body of a particular town. In such a situation, it would be convenient to select a 1-in-k systematic sample from the list of registered voters in place of the usual simple random sample. All relevant formulas corresponding to the estimator of P can be obtained from the formulas for $\bar{y}_{sy}$, by taking $y_i=0$ if the i-th sample unit does not possess the specified attribute, and $y_i=1$ if it does. The estimator p_{sy} would thus be the average of the 0 and 1 values assigned to the units in the sample. Assuming that n_1 units in the systematic sample of n units possess the attribute under study, expressions for the estimator p_{sy} of the population proportion, variance $V(p_{sy})$, and the estimator of this variance can be obtained from (6.1) to (6.5). These are given as follows :

Estimator of proportion P :

$$p_{sy} = \frac{n_1}{n} \tag{6.12}$$

Variance of estimator p_{sy} :

$$V(p_{sy}) = \frac{1}{k} \sum_{r=1}^{k} (p_{sy} - P)_r^2 \qquad \text{(for LS sampling)} \tag{6.13}$$

$$= \frac{1}{N} \sum_{r=1}^{N} (p_{sy} - P)_r^2 \qquad \text{(for CS sampling)} \tag{6.14}$$

Estimator of variance $V(p_{sy})$:

$$v(p_{sy}) = \frac{(N-n)\,R}{2Nn\,(n-1)} \tag{6.15}$$

where R is the total number of times that 0 follows 1 or 1 follows 0 in the ordered sequence of observations for the n sample units.

Estimator of $V(p_{sy})$ for population in random order :

$$v(p_{sy}) = \frac{(N-n)}{N} \left[\frac{p_{sy}(1-p_{sy})}{(n-1)} \right] \tag{6.16}$$

In case one wishes to estimate the total number N_1 of units in the population that possess the desired attribute, the estimator $\hat{N}_1$ is obtained by multiplying p_{sy} by N. Also, the variance $V(\hat{N}_1)$ and its estimator $v(\hat{N}_1)$ are N^2 times the corresponding expressions for p_{sy}.

We now take up an example to illustrate the various steps involved in estimating proportion and the number of units in the population possessing the attribute under study.

Example 6.7

On public complaint that some gas cylinders supplied for domestic use were underweight, an inquiry committee was set up. The committee decided to examine 1-in-50 cylinders from the 8000 cylinders stored in a warehouse, arranged in rows by the gas company. The committee found 18 cylinders to be underweight from the 160 sampled cylinders. Estimate the total number N_1, and also the proportion, of underweight cylinders in the warehouse. Also, build up confidence interval for these parameters.

Solution

We have N= 8000, n=160, and n_1=18. Then the estimate of proportion of underweight cylinders in the warehouse is

$$p_{sy} = \frac{n_1}{n}$$
$$= \frac{18}{160}$$
$$= .1125$$

Estimated total number of underweight cylinders would then be

$$\hat{N}_1 = Np_{sy}$$
$$= (8000)(.1125)$$
$$= 900$$

Assuming that the population units (gas cylinders) under study were placed in random order before drawing the sample, the estimate of variance $V(\hat{N}_1)$ is computed using (6.16) as

$$v(\hat{N}_1) = N^2 v(p_{sy})$$
$$= \frac{N(N-n)\, p_{sy}(1-p_{sy})}{(n-1)}$$
$$= \frac{8000(8000-160)(.1125)(1-.1125)}{159}$$
$$= 39384.90$$

Also, one can work out the confidence interval for N_1 from

$$\hat{N}_1 \pm 2\sqrt{v(\hat{N}_1)}$$
$$= 900 \pm 2\sqrt{39384.90}$$
$$= 503.09,\ 1296.91$$
$$\approx 503,\ 1297$$

Thus, the inquiry committee estimated 900 underweight cylinders from this particular sample. The committee also feels that the total number of underweight cylinders in the warehouse is likely to be between 503 and 1297.

From the above calculations, we find that $p_{sy} = .1125$ is the estimate of the proportion of underweight cylinders in the warehouse. Also,

$$v(p_{sy}) = \frac{v(\hat{N}_1)}{N^2}$$

$$= \frac{39384.90}{(8000)^2}$$

$$= .0006154$$

The required confidence interval for the proportion of underweight cylinders is given by

$$p_{sy} \pm 2\sqrt{v(p_{sy})}$$

$$= .1125 \pm 2\sqrt{.0006154}$$

$$= .0629, .1621$$

Thus, the proportion of underweight cylinders in the warehouse is most probably in the range .0629 to .1621. ■

It can be noted here that the lower and upper limits for the confidence interval of P can alternatively be obtained by dividing corresponding limits for the confidence interval for N_1 by N.

6.7 SOME FURTHER REMARKS

6.1 In case of $N \neq nk$, the sample mean $\bar{y}_{sy}$ based on a LS sample does not remain unbiased for the population mean $\bar{Y}$. This problem can also be resolved if the procedure in selecting the random start is slightly modified. The modification consists of selecting random start r, $1 \leq r \leq k$, with probability $P(r)=n_r/N$, where n_r is the number of units to be selected in the systematic sample corresponding to the random start r. For example, if we have N=14, n=5, so that k = 3, then random starts r = 1, 2 will yield 5 units in each sample, while r = 3 will result in the selection of 4 units. Thus, if we associate with r = 1 and r = 2 a probability of 5/14 while r = 3 is selected with probability 4/14, the sample mean $\bar{y}_{sy}$ will become unbiased for the population mean.

6.2 In case of populations with *linear trend*, the relative efficiency of systematic sample mean is very high in relation to the simple random sample mean for estimating the mean of such populations. Consider a hypothetical population in which the model $Y_i=a+b.i$, where a and b are constants and i is the serial number of the unit in the frame with study variable value Y_i, exactly holds. For such a population, the relative efficiency is approximately equal to the sample size.

6.3 For populations with linear trend, Yates (1948) has made a suggestion known as *Yates end correction*, which helps in greatly reducing the error in the estimator $\bar{y}_{sy}$. We find that in estimator $\bar{y}_{sy}$ in (6.1), all the observations $(y_1, y_2,...,y_n)$ received a weight equal to 1/n. Yates has proposed certain corrections to the weights

associated with the end units (first and last) in the sample. In place of 1/n, the first unit receives a weight $\left(\frac{1}{n}+x\right)$ whereas the weight associated with the last unit equals $\left(\frac{1}{n}-x\right)$, where $x=\frac{2r-k-1}{2(n-1)k}$. Thus, the new estimator of population mean becomes

$$\bar{y}'_{sy} = \bar{y}_{sy} + \frac{2r-k-1}{2(n-1)k}(y_1 - y_n)$$

6.4 For the population with linear trend and N=nk, a sampling procedure known as *balanced systematic sampling*, is also helpful in reducing the error of the estimator $\bar{y}_{sy}$. The procedure assumes that the serial numbers of the units in the population are divided into n/2 groups of 2k (sampling interval) units each. A pair of units, equidistant from the end units of that group, is then systematically selected from each group. For example, if r is the random start selected from 1 to k, the units with serial numbers r and 2k-r+1 will be selected from the first group of 2k units. Then, the two units selected from the second group of 2k units, will be the units with serial numbers 2k+r and 4k-r+1, and so on. Thus, the balanced systematic sample of n (even) units with random start r will consist of units with serial numbers [r+2jk, 2(j+1)k-r+1], j=0,1,2,...,(n/2 - 1).

LET US DO

6.1 What is systematic sampling? Discuss its merits and demerits.

6.2 Differentiate between systematic sampling and simple random sampling. Do you think that in presence of periodicity in the population, systematic sampling can be used more efficiently ? If so, how could it be done ?

6.3 The circular systematic sampling is usually preferred over linear systematic sampling. Discuss, why is it so ?

6.4 The number of colleges in 12 districts of a state are 8, 10, 6, 7, 7, 9, 11, 5, 6, 8, 9, and 11. List all possible samples of size 3 that can be selected from this population of 12 units using LS and CS sampling. Also, determine the average of corresponding sample means in both the cases. Are the two averages equal to the population mean? If yes, what does it indicate about the bias in the two estimators ?

6.5 Many trees along a canal have been uprooted by a storm. This damage persists along a 35 km stretch. The Department of Irrigation is interested in estimating total number of these damaged trees. Each one kilometer segment along the canal has been divided into 5 equal parts by stone markers. Thus, the entire 35 km long stretch is divided into 175 equal segments. Twenty five of these segments are selected using LS sampling with a sampling interval of 7 segments. The information regarding number of uprooted trees (y) obtained from this 1-in-7 systematic sample is given in the following table :

Selected segment	y	Selected segment	y	Selected segment	y
6	4	62	3	118	23
13	17	69	8	125	12
20	11	76	5	132	8
27	6	83	13	139	17
34	8	90	9	146	6
41	16	97	16	153	5
48	21	104	17	160	8
55	13	111	9	167	10
				174	15

Estimate the total number of uprooted trees, and also determine the confidence interval for it.

6.6 It is desired to estimate the average per day rent for single occupancy rooms in well known hotels of a state. In all, there are 192 such hotels in the state and these are listed in a book entitled "A Guide to Visitors". The investigator selected a 1-in-8 sample of hotels and rang up the managers of sampled hotels. The information on rent (in rupees) so obtained is given below :

Hotel	Rent	Hotel	Rent	Hotel	Rent
1	100	9	90	17	125
2	120	10	110	18	85
3	125	11	125	19	90
4	115	12	80	20	105
5	110	13	70	21	130
6	80	14	125	22	95
7	130	15	130	23	135
8	120	16	105	24	140

Estimate the average per day rent along with the confidence limits for it.

6.7 The editor of a local daily newspaper is interested in estimating average number of misprints in the daily over a year of 365 days. All the editions of the daily corresponding to 365 days were labeled from 1 to 365. A systematic sample of 28 editions of the paper was selected using CS sampling, taking $k=365/28 \approx 13$. The selected editions were then carefully examined and misprints counted. These are given in the following table along with the serial numbers of the selected editions of the paper.

Edition No.	Number of misprints	Edition No.	Number of misprints	Edition No.	Number of misprints
73	3	190	9	307	11
86	11	203	4	320	9
99	4	216	8	333	6
112	2	229	6	346	10
125	8	242	5	359	8
138	0	255	7	7	1
151	6	268	16	20	5
164	13	281	10	33	0
177	5	294	8	46	3
				59	7

Estimate average number of misprints in a daily edition, and place confidence limits on it.

6.8 There are 280 wells in a region for recording water table depth. Owing to heavy rainfall in the area under study, the water table has gone up considerably and is likely to cause waterlogging in the area. For estimating the average water table depth, the irrigation department selected a 1-in-10 sample of wells, and made the following observations on the water table depth (y) in meters.

Well	y	Well	y	Well	y	Well	y
1	2.60	8	2.40	15	2.60	22	2.10
2	2.80	9	2.60	16	1.40	23	2.00
3	2.85	10	2.35	17	1.55	24	1.80
4	3.75	11	1.85	18	1.80	25	2.35
5	2.85	12	3.30	19	2.50	26	1.60
6	2.45	13	2.50	20	2.30	27	1.70
7	1.90	14	2.10	21	1.70	28	2.10

Estimate mean water table depth in the region, and place confidence limits on it.

6.9 A reputed public school has 800 children. The school management has changed the school uniform twice in a year. A sociologist wishes to gauge parents' reaction to this decision of the management. The sociologist has 5 skilled investigators. He selected 5 subsamples of 8 students each, so that, the overall sample was of size 40. Each investigator interviewed the parents of the students falling in the subsample assigned to him, and gave scores from 1 to 10 depending on the severity of respondent's reaction to the management's decision. The scores thus recorded are presented in the following table :

Subsample	Scores							
1	3	5	1	6	8	4	9	10
2	1	7	5	3	6	4	7	8
3	6	2	8	3	6	5	2	4
4	9	4	6	5	4	3	1	2
5	5	3	9	6	1	4	8	1

Assuming that the random starts for the subsamples were selected with replacement, estimate the average score for the parents of all the 800 students, and also build up the confidence interval for the true average score.

6.10 A graduate student in statistics was given an assignment to estimate average height of all the 900 graduate students at a certain university. The student selected an overall sample of 60 graduate students in the form of six subsamples, each consisting of 10 students, using systematic sampling. The average height (in cm) of students in each subsample was computed, and is given below :

$\bar{y}_1 = 160.8$ $\bar{y}_2 = 168.6$ $\bar{y}_3 = 169.4$
$\bar{y}_4 = 164.5$ $\bar{y}_5 = 163.4$ $\bar{y}_6 = 166.8$

Estimate the average height of all the 900 students, and construct confidence interval for it. Assume that the random starts for the subsamples were selected through SRS without replacement.

6.11 Assume that the sample of 24 hotels selected in exercise 6.6 is a preliminary sample. Examine whether this sample is sufficient to estimate the average per day rent with a permissible error of Rs 5 ? If not, how many additional units need to be selected.

6.12 Some of the school buildings in a district collapsed during last few years, and caused damage to life and property. The district administration decided to have a quick estimate of the proportion of unsafe school buildings in the district. For this purpose, a systematic sample of 84 buildings, out of a total of 1260 school buildings, was selected . The selected school buildings were examined by experts. The number of unsafe buildings was found to be 16. Estimate the proportion of unsafe buildings in the district, and work out the confidence interval for it.

6.13 District traffic police is concerned about the vehicle owners not carrying necessary documents with them. To estimate the seriousness of the problem, a check point was set up on the Grand Trunk Road. Due to heavy rush of traffic, every 20th vehicle was stopped and its papers examined. In all, 114 vehicles were checked, of which 19 were not having necessary documents. Estimate the proportion of vehicle owners who do not carry the required documents. Also, construct confidence interval for this proportion.

6.14 Consider a population of units exhibiting a linear trend. Would you prefer using systematic sampling, or the usual SRS, for estimating mean/total ? Give reasons in support of your decision.

CHAPTER 7

Ratio and Product Methods of Estimation

7.1 NEED FOR RATIO ESTIMATION

In the preceding chapters, we have discussed some methods of using information on an auxiliary variable for improving the precision of the estimates of population mean/total. In chapter 4, the selection probabilities for the population units were determined from the measures of size provided by such supplementary information. Also, the use of information on the auxiliary variable for the purpose of stratification has been discussed in chapter 5. In this chapter, and also in the following chapter, we present some other estimators that make use of auxiliary information for achieving higher efficiency.

In socio-economic surveys, one may be interested in estimating ratios like per capita income or expenditure. Similarly, estimation of yield per unit area, or the use of fertilizer/pesticides per hectare for a particular crop, could be of importance in case of agricultural surveys. Estimation of percent relative fall in real estate prices or input-output ratio, are useful for industry and commerce.

Population ratio R is the ratio of two population parameters. Mostly, these parameters are population totals or means. For instance, the yield per hectare R=Y/X, where Y is the total production and X the corresponding total area under the crop. The ratio R is usually estimated by the ratio of unbiased estimators $\hat{Y}$ and $\hat{X}$ of Y and X respectively. It could also be equivalently estimated by $\hat{R}=\bar{y}/\bar{x}$, where $\bar{y}$ and $\bar{x}$ unbiasedly estimate the population means $\bar{Y}$ and $\bar{X}$ respectively. Here x, the area under the crop, is treated as the auxiliary variable. The estimators of such population ratios are known as *ratio estimators*.

The estimators of population ratio R can also be used for building up the estimators of population mean/total for the study variable y. In situations, where the study variable y is highly correlated with the auxiliary variable x, and the two are also approximately proportional, the ratio of y to x is expected to be less variable than the y's themselves. In such a situation, it would, therefore, be better to estimate R from the sample and multiply the estimator of R with the known population mean/total of the auxiliary variable x, to obtain an estimator for the population mean/total of the study variable y. The estimator, so obtained, is also called ratio estimator of population mean $\bar{Y}$ or the total Y.

We first consider the estimator of population ratio R. The estimators for mean/total shall be discussed subsequently.

7.2 ESTIMATION OF POPULATION RATIO

Suppose that a WOR simple random sample of n units is drawn from a population of N units to estimate the population ratio

$$R = \frac{Y}{X} \tag{7.1}$$

where, as mentioned in the preceding section, $Y = \Sigma Y_i$ and $X = \Sigma X_i$, i = 1, 2, ..., N, are the population totals for the estimation variable y and the auxiliary variable x respectively. We assume that the population mean $\bar{X} = X/N$ is known. All the sample units are then observed for the variables y and x. Let $(y_1, x_1), (y_2, x_2),...,(y_n, x_n)$ denote the set of these observations, whereas $\bar{y}$ and $\bar{x}$ represent the corresponding sample means.

As defined earlier, the population mean squares and product are given by

$$\left.\begin{aligned} S_y^2 &= \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right) \\ S_x^2 &= \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i^2 - N\bar{X}^2\right) \\ S_{xy} &= \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i Y_i - N\bar{X}\bar{Y}\right) \end{aligned}\right] \tag{7.2}$$

Also, their respective sample estimators are

$$\left.\begin{aligned} s_y^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) \\ s_x^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \\ s_{xy} &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \end{aligned}\right] \tag{7.3}$$

Further, let ρ and r be the population and sample correlation coefficients between the variables y and x respectively. We then have :

Estimator of population ratio R :

$$\hat{R} = \frac{\bar{y}}{\bar{x}} \tag{7.4}$$

Approximate bias of estimator $\hat{R}$:

$$B(\hat{R}) = \left(\frac{N-n}{Nn}\right)\left(\frac{1}{\bar{X}^2}\right)(RS_x^2 - S_{xy}) \tag{7.5}$$

Approximate mean square error of estimator $\hat{R}$:

$$\text{MSE}(\hat{R}) = \left(\frac{N-n}{Nn}\right)\left(\frac{1}{\bar{X}^2}\right)(S_y^2 + R^2 S_x^2 - 2RS_{xy}) \tag{7.6}$$

Estimator of MSE ($\hat{R}$) :

$$\text{mse}(\hat{R}) = \left(\frac{N-n}{Nn}\right)\left(\frac{1}{\bar{X}^2}\right)(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{xy}) \tag{7.7}$$

In case $\bar{X}$ is not known, the sample mean $\bar{x}$ could be used in its place in (7.7) to find the value of mse ($\hat{R}$).

Example 7.1

A survey project was undertaken by a graduate student in a small town of USA consisting of N=1620 family households. The purpose of the survey was to estimate the proportion of monthly family income spent on purchasing milk. An SRS without replacement sample of n=30 households was selected for this purpose. The sample data, in respect of monthly income (x) and expenditure on milk (y), both in dollars, are listed in table 7.1.

Table 7.1 Monthly family income and the amount spent on milk

Household	Monthly income	Expenditure on milk	Household	Monthly income	Expenditure on milk
1	2000	60	16	1620	60
2	2900	85	17	1880	48
3	2400	80	18	2402	67
4	3200	106	19	2665	68
5	2400	60	20	1948	62
6	3260	84	21	1870	50
7	1600	45	22	3400	94
8	3080	106	23	1700	53
9	2445	75	24	1805	50
10	3600	102	25	2850	100
11	2960	75	26	3700	108
12	1750	60	27	2960	84
13	1980	60	28	2730	54
14	4260	108	29	2840	66
15	3370	91	30	2620	52

Taking $\bar{X}$ = \$2550, find both point and interval estimates of R.

Solution

First of all, we work out the following statistics :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$= \frac{1}{30} (2000 + 2900 + ... + 2620)$$

$$= 2606.5$$

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

$$= \frac{1}{30} (60 + 85 + ... + 52)$$

$$= 73.77$$

$$s_x^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)$$

$$= \frac{1}{30-1} [(2000)^2 + (2900)^2 + ... + (2620)^2 - 30 (2606.5)^2]$$

$$= 489759.15$$

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

$$= \frac{1}{30-1} [(60)^2 + (85)^2 + ... + (52)^2 - 30(73.77)^2]$$

$$= 417.88$$

$$s_{xy} = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \right)$$

$$= \frac{1}{30-1} [(2000)(60) + (2900)(85) + ... + (2620)(52)$$

$$- 30 (2606.5) (73.77)]$$

$$= \frac{1}{29} [6125305 - 30 (2606.5) (73.77)]$$

$$= 12305.51$$

We now compute estimate of ratio R from (7.4) as

$$\hat{R} = \frac{\bar{y}}{\bar{x}}$$

$$= \frac{73.77}{2606.5}$$

$$= .02830$$

The estimate of mean square error for $\hat{R}$ is provided by (7.7). Using the sample values computed above, we have

$$\text{mse}(\hat{R}) = \left(\frac{N-n}{Nn}\right)\left(\frac{1}{\bar{X}^2}\right)(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{xy})$$

$$= \left[\frac{1620-30}{(1620)(30)}\right]\left[\frac{1}{(2550)^2}\right][417.88 + (.02830)^2 (489759.15)$$

$$- 2(.02830)(12305.51)]$$

$$= 5.71713 \times 10^{-7}$$

Now, we proceed to work out confidence interval for the population ratio R. It can be computed from

$$\hat{R} \pm 2\sqrt{\text{mse}(\hat{R})}$$

$$= .02830 \pm 2\sqrt{5.71713 \times 10^{-7}}$$

$$= .02830 \pm .00151$$

$$= .02679, .02981$$

Thus the proportion of monthly income spent on milk is estimated to be 2.830%, and the confidence limits above indicate that the population proportion is most likely to be in the range of 2.679% to 2.981%. ■

7.3 RATIO ESTIMATOR FOR POPULATION MEAN/ TOTAL

As mentioned earlier, the estimator of population mean $\bar{Y}$ can be obtained by multiplying the estimator $\hat{R}$ by $\bar{X}$, the population mean for auxiliary variable x. The expressions for bias, mean square error, and the estimator of mean square error for the estimator of mean can also be obtained by multiplying the expression for bias $B(\hat{R})$ in (7.5) by $\bar{X}$, and the expressions for MSE $(\hat{R})$ and mse $(\hat{R})$ in (7.6) and (7.7) by $\bar{X}^2$. This yields the following results :

Ratio estimator of population mean $\bar{Y}$:

$$\bar{y}_r = \frac{\bar{y}\,\bar{X}}{\bar{x}} \tag{7.8}$$

Approximate bias of estimator $\bar{y}_r$:

$$B(\bar{y}_r) = \left(\frac{N-n}{Nn}\right)\left(\frac{1}{\bar{X}}\right)(RS_x^2 - S_{xy})$$

$$= \left(\frac{N-n}{Nn}\right)\bar{Y}(C_x^2 - \rho C_x C_y) \tag{7.9}$$

Approximate mean square error of estimator $\bar{y}_r$:

$$MSE(\bar{y}_r) = \frac{N-n}{Nn}(S_y^2 + R^2 S_x^2 - 2RS_{xy}) = \left(\frac{N-n}{Nn}\right)\bar{Y}^2(C_y^2 + C_x^2 - 2\rho C_x C_y) \quad (7.10)$$

Estimator of MSE $(\bar{y}_r)$:

$$mse(\bar{y}_r) = \frac{N-n}{Nn}(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{xy}) \quad (7.11)$$

where C_y and C_x are the population coefficients of variation for the variables y and x respectively.

The expressions relating to estimator of population total can be easily obtained from the expressions for the mean.

Ratio estimator of population total Y :

$$\hat{Y}_r = N\bar{y}_r = \frac{\bar{y}X}{\bar{x}} \quad (7.12)$$

Bias of estimator $\hat{Y}_r$:

$$B(\hat{Y}_r) = N B(\bar{y}_r) \quad (7.13)$$

Mean square error of estimator $\hat{Y}_r$:

$$MSE(\hat{Y}_r) = N^2 MSE(\bar{y}_r) \quad (7.14)$$

Estimator of MSE $(\hat{Y}_r)$:

$$mse(\hat{Y}_r) = N^2 mse(\bar{y}_r) \quad (7.15)$$

The $MSE(\bar{y}_r)$ and $MSE(\hat{Y}_r)$ given in (7.10) and (7.14), will be smaller than their respective counterparts $V(\bar{y})$ and $V(\hat{Y})$ in (3.9) and (3.12), for the usual estimators based on simple random sampling, when

$$\rho > \frac{C_x}{2C_y} \quad (7.16)$$

If x is the same character as y but has been measured on an earlier occasion, the coefficients of variation C_x and C_y may be taken as equal. In that case, it pays to use the ratio method of estimation in place of simple mean estimator if $\rho > .5$. However, one should keep in mind that the inequality (7.16) is based on approximation. It should also be noted that the ratio estimate may not be as good as the simple average, even in the presence of perfect correlation between y and x, when the regression line of y on

x passes through a point on y-axis that is far from origin, since the near proportionality between y and x does not exist in that situation. To summarize :

> The ratio estimators of population mean and total will be more efficient than the usual SRS based respective estimators if $\rho > C_x / 2C_y$ and the regression line passes through, or nearly through, the origin.

It would be a sound practice to examine the relationship between y and x on the basis of past surveys, and use this information during future studies.

Example 7.2

The data on study variable (y) and auxiliary variable (x) given below, are for a hypothetical population of 8 units :

y :	10	12	15	17	18	22	24	30
x :	4	6	9	10	13	14	16	20

Work out the efficiency of ratio estimator $\bar{y}_r$ in relation to the usual estimator $\bar{y}$ for WOR simple random samples of size 3.

Solution

For calculating the desired relative efficiency, we shall need the values of R, S_y^2, S_x^2, and S_{xy}. So, we first obtain these values from the population data. Thus,

$$\bar{Y} = \frac{1}{8}(10+12+\ldots+30)$$

$$= 18.5$$

$$\bar{X} = \frac{1}{8}(4+6+\ldots+20)$$

$$= 11.5$$

$$R = \frac{\bar{Y}}{\bar{X}} = \frac{18.5}{11.5} = 1.6087$$

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

$$= \frac{1}{8-1}\left[(10)^2 + (12)^2 + \ldots + (30)^2 - 8(18.5)^2\right]$$

$$= \frac{1}{7}\left[3042 - 8(18.5)^2\right]$$

$$= 43.4286$$

$$S_x^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i^2 - N\bar{X}^2\right)$$

$$= \frac{1}{8-1}\left[(4)^2 + (6)^2 + \ldots + (20)^2 - 8(11.5)^2\right]$$

$$= \frac{1}{7}\left[1254 - 8(11.5)^2\right]$$

$$= 28.0000$$

$$S_{xy} = \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i Y_i - N\bar{X}\,\bar{Y}\right)$$

$$= \frac{1}{8-1}\left[(4)(10) + (6)(12) + \ldots + (20)(30) - 8(11.5)(18.5)\right]$$

$$= \frac{1}{7}\left[1943 - 8(11.5)(18.5)\right]$$

$$= 34.4286$$

The correlation coefficient between variables y and x is equal to

$$\rho = \frac{S_{xy}}{S_x S_y}$$

$$= \frac{34.4286}{\sqrt{(28.0000)(43.4286)}}$$

$$= .9873$$

On using above computed values in (3.9) and (7.10), we obtain variance $V(\bar{y})$ and mean square error $MSE(\bar{y}_r)$. Thus,

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

$$= \frac{8-3}{(8)(3)}(43.4286)$$

$$= 9.0476$$

$$MSE(\bar{y}_r) = \frac{N-n}{Nn}\left(S_y^2 + R^2 S_x^2 - 2RS_{xy}\right)$$

$$= \frac{8-3}{(8)(3)}\left[43.4286 + (1.6087)^2(28.0000) - 2(1.6087)(34.4286)\right]$$

$$= 1.0666$$

The $MSE(\bar{y}_r)$ is seen to be less than $V(\bar{y})$, implying that, the ratio estimator of population mean is more efficient than the usual SRS based estimator $\bar{y}$. The percent relative efficiency of the estimator $\bar{y}_r$ with respect to $\bar{y}$ is obtained as

$$RE = \frac{V(\bar{y})}{MSE(\bar{y}_r)}(100)$$

$$= \frac{9.0476}{1.0666}(100)$$

$$= 848.27 \blacksquare$$

Example 7.3

The agricultural wing of a district administration wishes to estimate the area under paddy harvested with combine. The district is comprised of 520 villages, including small towns, and the total area under paddy is 36,000 hectares. Keeping in view the budget at disposal, a WOR simple random sample of 26 villages was drawn for the purpose. The information on area in hectares, collected in respect of these villages, is given below in table 7.2.

Table 7.2 Area under paddy (x) and the area harvested with combine (y)

Village	x	y	Village	x	y
1	120	82	14	71	56
2	47	18	15	57	41
3	62	47	16	64	52
4	90	77	17	49	30
5	83	67	18	68	43
6	106	98	19	56	32
7	52	36	20	54	27
8	57	37	21	81	61
9	81	58	22	66	39
10	52	37	23	112	87
11	66	52	24	86	61
12	78	68	25	48	22
13	41	15	26	67	35

Estimate the total area under paddy in the district, harvested with combine, and obtain the confidence limits for this area.

Solution

We have X=36,000. As in example 7.1, we first work out the following sample estimates :

$$\bar{y} = \frac{1}{26}(82 + 18 + ... + 35)$$

$$= 49.15$$

$$\bar{x} = \frac{1}{26}(120 + 47 + ... + 67)$$

$$= 69.77$$

$$s_y^2 = \frac{1}{26-1}\left[(82)^2 + (18)^2 + \ldots + (35)^2 - 26(49.15)^2\right]$$

$$= 467.02$$

$$s_x^2 = \frac{1}{26-1}\left[(120)^2 + (47)^2 + \ldots + (67)^2 - 26(69.77)^2\right]$$

$$= 422.74$$

$$s_{xy} = \frac{1}{26-1}\left[(120)(82) + (47)(18) + \ldots + (67)(35) - 26(69.77)(49.15)\right]$$

$$= \frac{1}{25}\left[99624 - 26(69.77)(49.15)\right]$$

$$= 418.60$$

Using (7.12), we now estimate the total area under paddy harvested with combine. It gives

$$\hat{Y}_r = \frac{\bar{y}\,X}{\bar{x}}$$

$$= \frac{(49.15)(36000)}{69.77}$$

$$= 25360.47$$

Also,

$$\hat{R} = \frac{49.15}{69.77} = .7045$$

Estimate of mean square error can be obtained from (7.15) and (7.11). Thus,

$$\text{mse}(\hat{Y}_r) = N^2\,\text{mse}(\bar{y}_r)$$

$$= \frac{N(N-n)}{n}\left(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}\,s_{xy}\right)$$

$$= \frac{520(520-26)}{26}\left[467.02 + (.7045)^2(422.74) - 2(.7045)(418.60)\right]$$

$$= 859826.98$$

Following (2.8), the confidence limits for population total are obtained as

$$\hat{Y}_r \pm 2\sqrt{\text{mse}(\hat{Y}_r)}$$

$$= 25360.47 \pm 1854.54$$

$$= 23505.93,\ 27215.01$$

It means, the total area under paddy, harvested with combine, in the district is likely to fall in the closed interval [23505.93, 27215.01] hectares with probability approximately equal to .95.

It must be pointed out here, that if the information on the auxiliary variable (area under paddy) was not used in the form of ratio estimator to estimate the total paddy area harvested with combine, the other alternative was to estimate it through the estimator $\hat{Y} = N\bar{y}$, where $\bar{y}$ is the usual mean based on WOR simple random sample. The estimator of the variance $V(\hat{Y})$ would then have been, from (3.10) and (3.13), as

$$\begin{aligned} v(\hat{Y}) &= \frac{N(N-n)}{n} s_y^2 \\ &= \frac{520(520-26)}{26} (467.02) \\ &= 4614157.60 \end{aligned}$$

which is much larger than $mse(\hat{Y}_r)$. Estimated percent relative efficiency of the ratio estimator, with respect to the SRS without replacement estimator $\hat{Y}$, is given by

$$\begin{aligned} RE &= \frac{v(\hat{Y})}{mse\,(\hat{Y}_r)} (100) \\ &= \frac{4614157.60}{859826.98} (100) \\ &= 536.64 \end{aligned}$$

Thus the ratio estimator $\hat{Y}_r$ for the total combine harvested area is over five times more efficient than the without replacement SRS estimator. This increase in efficiency is due to the use of the auxiliary information on the area under paddy. ■

7.4 DETERMINING THE SAMPLE SIZE FOR ESTIMATION OF RATIO, MEAN, AND TOTAL

Once the sampling design for the study has been chosen, the next question that the investigator faces is to decide about the number of units to be selected in the sample so as to get estimators with predetermined precision. In this section, we discuss the procedure to determine the required sample size for estimating population parameters R, $\bar{Y}$, and Y with B as the bound on error of estimation.

To determine the number of sample units required to estimate the population ratio R, defined in (7.1), with B as the bound on the estimation error, we need to solve the equation

$$2\sqrt{mse\,(\hat{R})} = B \qquad (7.17)$$

for n. The expression for mse($\hat{R}$) has been given in (7.7). For evaluating the value of mse($\hat{R}$), we make use of the observations on the variables y and x for a preliminary sample of size n_1. Thus,

$$\text{mse}(\hat{R}) = \left(\frac{N-n}{Nn}\right)\frac{1}{\bar{X}^2}(s_{y1}^2 + \hat{R}_1^2 s_{x1}^2 - 2\hat{R}_1 s_{xy1})$$

where the sample mean squares and product s_{y1}^2, s_{x1}^2, and s_{xy1} and the ratio $\hat{R}_1$ are obtained from the preliminary sample of size n_1. The solution of (7.17) for n, gives the following rule for choosing the required sample size.

Sample size to estimate R with a bound B on error of estimation :

$$n = \frac{Ns_r^2}{ND + s_r^2} \tag{7.18}$$

where

$$s_r^2 = s_{y1}^2 + \hat{R}_1^2 s_{x1}^2 - 2\hat{R}_1 s_{xy1} \tag{7.19}$$

and

$$D = \frac{\bar{X}^2 B^2}{4}$$

If $n_1 \geq n$, the preliminary sample of size n_1 is enough. Include $(n-n_1)$ additional units in the sample, otherwise.

Example 7.4

Assuming the sample of 30 houses drawn in example 7.1 as a preliminary sample, determine the sample size required for estimating ratio R of example 7.1 with a margin of error .001.

Solution

Since the sample of 30 houses of example 7.1 is now taken as the preliminary sample, the sample estimates worked out in example 7.1 will be treated as estimates obtained from the preliminary sample. This means,

$$n_1 = 30,\ \bar{x} = \bar{x}_1 = 2606.5,\ \hat{R} = \hat{R}_1 = .02830,\ s_y^2 = s_{y1}^2 = 417.88,$$

$$s_x^2 = s_{x1}^2 = 489759.15, \text{ and } s_{xy} = s_{xy1} = 12305.51.$$

From (7.19), we then work out

$$\begin{aligned} s_r^2 &= s_{y1}^2 + \hat{R}_1^2 s_{x1}^2 - 2\hat{R}_1 s_{xy1} \\ &= 417.88 + (.02830)^2 (489759.15) - 2\,(.02830)\,(12305.51) \\ &= 113.6313 \end{aligned}$$

Also,

$$D = \frac{\bar{X}^2 B^2}{4}$$

$$= \frac{(2550)^2 (.001)^2}{4}$$

$$= 1.6256$$

From (7.18), the sample size needed to estimate population ratio R, with a bound on the error of estimation as .001, would be

$$n = \frac{Ns_r^2}{ND + s_r^2}$$

$$= \frac{(1620)(113.6313)}{(1620)(1.6256) + 113.6313}$$

$$= 67.01$$

$$\approx 67$$

This means that the investigator needs to select $n\text{-}n_1 = 67\text{-}30 = 37$ more households to estimate the population ratio R with the desired precision. ■

Similarly, when ratio estimator is used for estimating the population mean/total, the required sample size is obtained by solving the following equations for n :

$$2\sqrt{\text{mse}(\bar{y}_r)} = B \quad \text{(when estimating mean)}$$

$$2\sqrt{\text{mse}(\hat{Y}_r)} = B \quad \text{(when estimating total)}$$

where, as in (7.17), the value of $(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy})$ involved in $\text{mse}(\bar{y}_r)$ and $\text{mse}(\hat{Y}_r)$, defined in (7.11) and (7.15), is calculated from the n_1 observations on variables y and x from the preliminary sample, and B is the magnitude of bound on the error of estimation. This yields the following rule :

Sample size to estimate the population mean / total with a permissible error B :

$$n = \frac{Ns_r^2}{ND + s_r^2} \qquad (7.20)$$

where

$D = B^2/4$ (when estimating mean)

$D = B^2/4N^2$ (when estimating total)

and s_r^2 is as defined in (7.19). If $n_1 \geq n$, the sample size n_1 is sufficient, otherwise, one needs to select $(n\text{-}n_1)$ additional units in the sample.

Example 7.5

Treating the sample of 26 villages drawn in example 7.3 as the preliminary sample, determine the sample size required to estimate the total area under paddy, harvested with combine, with the bound on the error of estimation as 1500 hectares.

Solution

As the sample of 26 villages selected in example 7.3 is assumed as a preliminary sample, the sample estimates worked out there would be treated as estimates provided by the preliminary sample. We thus have

$$n_1 = 26,\ \hat{R} = \hat{R}_1 = .7045,\ s_y^2 = s_{y1}^2 = 467.02,\ s_x^2 = s_{x1}^2 = 422.74,$$

and $s_{xy} = s_{xy1} = 418.60$.

Then we compute

$$\begin{aligned} s_r^2 &= s_{y1}^2 + \hat{R}_1^2 s_{x1}^2 - 2\hat{R}_1 s_{xy1} \\ &= 467.02 + (.7045)^2 (422.74) - 2(.7045)(418.60) \\ &= 87.0270 \end{aligned}$$

Further,

$$D = \frac{B^2}{4N^2} = \frac{(1500)^2}{4(520)^2} = 2.0803$$

The required sample size can then be computed by using (7.20). Thus,

$$\begin{aligned} n &= \frac{Ns_r^2}{ND + s_r^2} \\ &= \frac{(520)(87.0270)}{(520)(2.0803) + 87.0270} \\ &= 38.7 \\ &\approx 39 \end{aligned}$$

This means that the investigator will be required to select 39-26=13 more villages, if the population total under study has to be estimated with a maximum margin of error as 1500 hectares. ■

7.5 SEPARATE AND COMBINED RATIO ESTIMATORS

For the reasons mentioned in chapter 5, it is sometimes desirable to stratify the population and then use ratio estimators for estimating population mean or total. For the discussion in this section, we shall assume that the sample drawn from each stratum is large enough, so that, the mean square approximations work fairly well.

Let us assume that the population of N units is divided into L strata, such that, h-th stratum has N_h units. Thus $\Sigma N_h = N$, h = 1,2,...,L. From the h-th stratum, a WOR

simple random sample of n_h units is selected, so that, the total sample size over all the strata becomes n. Also, let $\bar{Y}_h$ and $\bar{X}_h$, h=1,2,...,L, denote the h-th stratum means for the variables y and x respectively, whereas $\bar{y}_h$ and $\bar{x}_h$ denote their sample counterparts. The h-th stratum mean squares and product for the two variables are defined as

$$\left.\begin{aligned} S_{hy}^2 &= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 \\ S_{hx}^2 &= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2 \\ S_{hxy} &= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)(Y_{hi} - \bar{Y}_h) \end{aligned}\right] \qquad (7.21)$$

Similarly, s_{hy}^2, s_{hx}^2, and s_{hxy} are the sample mean squares and product which unbiasedly estimate S_{hy}^2, S_{hx}^2, and S_{hxy} respectively.

Now we discuss two different methods for constructing ratio estimators in stratified sampling. One is to build up the ratio estimators $(\bar{y}_h/\bar{x}_h)\bar{X}_h$, h = 1,2,...,L, within each stratum, and then form a weighted average of these separate estimators as a single estimator of population mean. The estimator, so obtained, is known as *separate ratio estimator*. Alternatively, one can estimate population mean $\bar{Y}$ by $\bar{y}_{st}$ and the mean $\bar{X}$ by $\bar{x}_{st}$ using stratified sampling. Then $(\bar{y}_{st}/\bar{x}_{st})\bar{X}$ can be used as an estimator of population mean $\bar{Y}$. This estimator is known as *combined ratio estimator*, and was proposed by Hansen *et al.* (1946).

Separate ratio estimator of population mean $\bar{Y}$:

$$\bar{y}_{sr} = \sum_{h=1}^{L} \frac{W_h \bar{y}_h}{\bar{x}_h} \bar{X}_h \qquad (7.22)$$

Approximate bias of the estimator $\bar{y}_{sr}$:

$$B(\bar{y}_{sr}) = \sum_{h=1}^{L} W_h \left(\frac{N_h - n_h}{N_h n_h}\right)\left(\frac{1}{\bar{X}_h}\right)(R_h S_{hx}^2 - S_{hxy}) \qquad (7.23)$$

Approximate mean square error of estimator $\bar{y}_{sr}$:

$$MSE(\bar{y}_{sr}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right)(S_{hy}^2 + R_h^2 S_{hx}^2 - 2R_h S_{hxy}) \qquad (7.24)$$

Estimator of MSE($\bar{y}_{sr}$) :

$$mse(\bar{y}_{sr}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right)(s_{hy}^2 + \hat{R}_h^2 s_{hx}^2 - 2\hat{R}_h s_{hxy}) \qquad (7.25)$$

where $\hat{R}_h = \bar{y}_h/\bar{x}_h$.

Example 7.6

A farm owner wishes to conduct a survey to estimate per tree yield for orange variety Mandarins (*Citrus reticulata*) in his orchard. The orchard has 8 rows of 30 trees each. The trees in the first 6 rows were planted 10 years back, and are now fully developed and matured. The last 2 rows consist of younger trees that were planted 6 years back. Record of yield for the preceding year for all the trees is available. Per tree yield for the matured plants was 98 kg, and it was 53 kg for the younger plants. In order to estimate current average yield, a stratified WOR simple random sample of 24 trees was selected using proportional allocation. Out of 24 trees in the sample, n_1=18 trees were selected from the first six rows (stratum I), and n_2= 6 trees came from the last two rows of younger trees (stratum II). All the selected trees were observed for yield. The yield (in kg) figures for the preceding year (x) and the current year (y) are given below :

Table 7.3 Orange yields for sample trees

Stratum I				Stratum II			
x	y	x	y	x	y	x	y
90.5	94.5	110.2	116.1	103.7	108.5	50.7	66.0
84.0	88.8	98.6	102.4	90.4	96.4	55.9	58.0
92.0	90.0	105.9	100.3	115.5	121.0	67.0	76.4
76.5	82.0	83.1	90.9	95.1	105.0	56.5	72.8
105.3	109.4	74.0	82.6	100.4	102.3	58.4	77.8
85.6	90.0	107.5	114.5	80.0	87.1	62.3	72.6

Estimate average yield per tree, using separate ratio estimator, and place confidence limits on it.

Solution

From the statement of the example, we have N = 240, N_1 = 180, N_2 = 60, n = 24, n_1 = 18, n_2 = 6, $\bar{X}_1$ = 98, and $\bar{X}_2$ = 53. The sample means, sample mean squares and products for the two strata have been computed, and are given below in table 7.4 along with certain other sample and population characteristics.

Table 7.4 Certain calculated strata values

Stratum I				Stratum II			
n_1 =	18	s^2_{1x} =	150.134	n_2 =	6	s^2_{2x} =	31.659
N_1 =	180	s^2_{1y} =	137.261	N_2 =	60	s^2_{2y} =	54.848
W_1 =	.75	s_{1xy} =	136.985	W_2 =	.25	s_{2xy} =	24.056
$\bar{X}_1$ =	98	r_1 =	.954	$\bar{X}_2$ =	53	r_2 =	.577
$\bar{x}_1$ =	94.350	$\hat{R}_1$ =	1.049	$\bar{x}_2$ =	58.467	$\hat{R}_2$ =	1.208
$\bar{y}_1$ =	98.989			$\bar{y}_2$ =	70.600		

The separate ratio estimate given by (7.22) yields

$$\begin{aligned}\bar{y}_{sr} &= \frac{W_1\bar{y}_1\bar{X}_1}{\bar{x}_1} + \frac{W_2\bar{y}_2\bar{X}_2}{\bar{x}_2} \\ &= W_1\hat{R}_1\bar{X}_1 + W_2\hat{R}_2\bar{X}_2 \\ &= (.75)(1.049)(98) + (.25)(1.208)(53) \\ &= 93.108\end{aligned}$$

We now work out mean square error of estimator $\bar{y}_{sr}$. Thus, from (7.25)

$$\begin{aligned}\text{mse}(\bar{y}_{sr}) &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_n\, n_h}\right)(s_{hy}^2 + \hat{R}_h^2\, s_{hx}^2 - 2\hat{R}_h\, s_{hxy}) \\ &= (.75)^2\left[\frac{180-18}{(180)(18)}\right][137.261 + (1.049)^2(150.134) \\ &\quad - 2(1.049)(136.985)] + (.25)^2\left[\frac{60-6}{(60)(6)}\right][54.848 \\ &\quad + (1.208)^2(31.659) - 2(1.208)(24.056)] \\ &= .4240 + .4024 \\ &= .8264\end{aligned}$$

The next step is to compute confidence limits. This we do through (2.8) as

$$\begin{aligned}&\bar{y}_{sr} \pm 2\sqrt{\text{mse}(\bar{y}_{sr})} \\ &= 93.108 \pm 1.818 \\ &= 91.290,\ 94.926\end{aligned}$$

Hence, the average orange yield per tree is estimated as 93.108 kg. Also, it is indicated that had all the trees in the orchard been observed for yield, the per tree yield would most probably have taken a value in the closed interval [91.290, 94.926]. ■

Unless the ratio R_h is constant from stratum to stratum, the separate ratio estimator is likely to be more precise than the combined estimator. For the separate ratio estimator, the sample size in all the strata should be sufficiently large, otherwise, it will have appreciable bias and the MSE($\bar{y}_{sr}$) approximations will not be good enough. It also needs the knowledge of $\bar{X}_h$ for each stratum. With only a small sample in each stratum, the combined estimator is to be recommended unless there is good empirical evidence to indicate wide differences in the strata ratios. Various expressions corresponding to the *combined ratio estimator* are given in the following box :

Combined ratio estimator of mean $\bar{Y}$:

$$\bar{y}_{cr} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X} = \left[\frac{\sum_{h=1}^{L} N_h \bar{y}_h}{\sum_{h=1}^{L} N_h \bar{x}_h} \right] \bar{X} \tag{7.26}$$

Approximate bias of the estimator $\bar{y}_{cr}$:

$$B(\bar{y}_{cr}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h} \right) \left(\frac{1}{\bar{X}} \right) (RS_{hx}^2 - S_{hxy}) \tag{7.27}$$

Approximate mean square error of estimator $\bar{y}_{cr}$:

$$MSE(\bar{y}_{cr}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_n n_h} \right) (S_{hy}^2 + R^2 S_{hx}^2 - 2RS_{hxy}) \tag{7.28}$$

Estimator of MSE($\bar{y}_{cr}$) :

$$mse(\bar{y}_{cr}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_n n_h} \right) (s_{hy}^2 + \hat{R}^2 s_{hx}^2 - 2\hat{R} s_{hxy}) \tag{7.29}$$

where $\hat{R} = \bar{y}_{st}/\bar{x}_{st}$.

Example 7.7

Since the ratios $\hat{R}_1$ and $\hat{R}_2$ computed in example 7.6 do not differ much, one can also use combined ratio estimator in place of separate ratio estimator. Thus, estimate the per tree yield using the estimator $\bar{y}_{cr}$, and also obtain the confidence interval for it.

Solution

Most of the intermediate values required for the purpose of estimation are already available in table 7.4. Hence, we have

$$\bar{y}_{st} = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h$$

$$= (.75)(98.989) + (.25)(70.600)$$

$$= 91.892$$

$$\bar{x}_{st} = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{x}_h = \sum_{h=1}^{L} W_h \bar{x}_h$$

$$= (.75)(94.350) + (.25)(58.467)$$

$$= 85.379$$

Also, the population mean $\bar{X}$ for all the 240 trees is calculated from the relation

$$\bar{X} = \sum_{h=1}^{L} W_h \bar{X}_h$$

$$= (.75)(98) + (.25)(53)$$

$$= 86.75$$

The combined ratio estimator of the average yield per tree, from (7.26), would be

$$\bar{y}_{cr} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X}$$

$$= \frac{(91.892)(86.75)}{85.379}$$

$$= 93.368$$

For computing estimated mean square error of $\bar{y}_{cr}$, we are to use single pooled value $\hat{R}$ in place of different $\hat{R}_1$ and $\hat{R}_2$ values for the two strata. This pooled $\hat{R}$ value is obtained from combined estimators $\bar{y}_{st}$ and $\bar{x}_{st}$. Thus,

$$\hat{R} = \frac{\bar{y}_{st}}{\bar{x}_{st}}$$

$$= \frac{91.892}{85.379}$$

$$= 1.076$$

We now work out estimated mean square error of $\bar{y}_{cr}$. From (7.29),

$$\text{mse}(\bar{y}_{cr}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right)(s_{hy}^2 + \hat{R}^2 s_{hx}^2 - 2\hat{R} s_{hxy})$$

Using above computed $\hat{R}$ and other values from example 7.6, one gets

$$\text{mse}(\bar{y}_{cr}) = (.75)^2 \left[\frac{180-18}{(180)(18)}\right][137.261 + (1.076)^2 (150.134)$$

$$- 2(1.076)(136.985)] + (.25)^2 \left[\frac{60-6}{(60)(6)}\right][54.848$$

$$+ (1.076)^2 (31.659) - 2(1.076)(24.056)]$$

$$= .4582 + .3725$$

$$= .8307$$

To work out the confidence interval we again use (2.8). Thus, we have the interval limits as

$$\bar{y}_{cr} \pm 2\sqrt{mse(\bar{y}_{cr})}$$

$$= 93.368 \pm 2\sqrt{.8307}$$

$$= 93.368 \pm 1.823$$

$$= 91.545,\ 95.191$$

The use of combined ratio estimator indicates that the population mean yield, based on all the 240 units, would most probably take a value in the range of 91.545 to 95.191 kg. ■

7.6 SOME FURTHER REMARKS

7.1 Let a sample of n units be selected using SRS without replacement. Also, if $r_i = y_i/x_i$ is the ratio of y and x values for i-th sample unit, and $\bar{r} = (\Sigma r_i)/n$, i=1,2,..., n, then the estimator

$$\bar{y}_{hr} = \bar{r}\,\bar{X} + \frac{n}{N}\left(\frac{N-1}{n-1}\right)(\bar{y} - \bar{r}\,\bar{x}) \tag{7.30}$$

is unbiased for population mean $\bar{Y}$. The variance of this estimator, to the first order of approximation, is equal to the $MSE(\bar{y}_r)$. The estimator $\bar{y}_{hr}$ was proposed by Hartley and Ross (1954).

7.2 If the sample is selected using Sen-Midzuno's scheme (Sen, 1952; Midzuno, 1952), where the first unit in the sample is selected with probability proportional to the auxiliary variable value and the remaining (n-1) units are selected through simple random sampling WOR, the ratio estimator $\bar{y}_r$, defined in (7.8), becomes unbiased for the population mean.

7.3 Another technique used to obtain unbiased estimators is based on splitting at random the sample of size n into k groups, each of size m = n/k. Let

$$\bar{y}_r^{(j)} = \frac{\bar{y}'_j}{\bar{x}'_j}\,\bar{X} \tag{7.31}$$

where $\bar{y}'_j$ and $\bar{x}'_j$ are the sample means based on a sample of (n-m) units obtained by omitting the j-th group of m units. Then the estimator

$$\bar{y}_m = \frac{1}{k}\sum_{j=1}^{k} \bar{y}_m^{(j)} \tag{7.32}$$

where

$$\bar{y}_m^{(j)} = \bar{y}_r^{(j)} + \frac{(N-n+m)}{N}\,k\left(\bar{y} - \bar{y}_r^{(j)}\,\frac{\bar{x}}{\bar{X}}\right) \tag{7.33}$$

is unbiased for the population mean $\bar{Y}$. This estimator is due to Mickey (1959).

7.4 Several other workers including Murthy and Nanjamma (1959), Quenouille (1956), Beale (1962), and Tin (1965) have proposed procedures of obtaining almost unbiased ratio type estimators for the population mean $\bar{Y}$. Some generalized estimators of population mean have been proposed by Srivastava (1967) and Diana (1993). Swain (1964) provides theoretical details for ratio estimators based on systematic samples.

7.7 PRODUCT METHOD FOR ESTIMATING MEAN / TOTAL

In section 7.3, we have seen that the ratio method of estimation provides a more efficient estimator of population mean/total than the usual SRS estimators of mean/total provided the values for the variables y and x are nearly proportional, and the correlation between them is positive and high. This means that the ratio estimator $\bar{y}_r$ can not be used to improve upon the conventional estimator $\bar{y}$ in situations where ρ is negative. For such cases, we consider an estimator known as *product estimator.*

Product estimator of population mean $\bar{Y}$:

$$\bar{y}_p = \frac{\bar{y}\,\bar{x}}{\bar{X}} \tag{7.34}$$

Approximate bias of estimator $\bar{y}_p$:

$$\left.\begin{aligned} B(\bar{y}_p) &= \left(\frac{N-n}{Nn}\right)\frac{S_{xy}}{\bar{X}} \\ &= \left(\frac{N-n}{Nn}\right)\bar{Y}\,\rho\,C_x C_y \end{aligned}\right] \tag{7.35}$$

Approximate mean square error of estimator $\bar{y}_p$:

$$\left.\begin{aligned} MSE(\bar{y}_p) &= \frac{N-n}{Nn}\,(S_y^2 + R^2 S_x^2 + 2RS_{xy}) \\ &= \left(\frac{N-n}{Nn}\right)\bar{Y}^2\,(C_y^2 + C_x^2 + 2\rho C_x C_y) \end{aligned}\right] \tag{7.36}$$

Estimator of MSE $(\bar{y}_p)$:

$$mse(\bar{y}_p) = \frac{N-n}{Nn}\,(s_y^2 + \hat{R}^2 s_x^2 + 2\hat{R}s_{xy}) \tag{7.37}$$

where C_y and C_x are the population coefficients of variation for the variables y and x respectively.

As usual, the expressions for the estimator of population total, its bias, mean square error, and estimator of mean square error can be written from the expressions for the estimator $\bar{y}_p$.

Estimator of population total Y:

$$\hat{Y}_p = N\bar{y}_p = \frac{N\bar{y}\,\bar{x}}{\bar{X}} \tag{7.38}$$

Bias of estimator $\hat{Y}_p$:

$$B(\hat{Y}_p) = N\,B(\bar{y}_p) \tag{7.39}$$

Mean square error of estimator $\hat{Y}_p$:

$$MSE(\hat{Y}_p) = N^2\,MSE(\bar{y}_p) \tag{7.40}$$

Estimator of MSE $(\hat{Y}_p)$:

$$mse(\hat{Y}_p) = N^2\,mse(\bar{y}_p) \tag{7.41}$$

The $MSE(\bar{y}_p)$ and $MSE(\hat{Y}_p)$, given in (7.36) and (7.40), will be smaller than the corresponding variances $V(\bar{y})$ and $V(\hat{Y})$, in (3.9) and (3.12) for the conventional estimators in case of SRS, if

$$\rho \leq -C_x/2C_y \tag{7.42}$$

implying $\rho \leq -.5$ when $C_x = C_y$. This yields the following statement :

The product estimator of population mean or total will be more efficient than the respective conventional estimator of mean or total in case of SRS, if $\rho \leq -C_x/2C_y$, which means $\rho \leq -.5$ when $C_x = C_y$.

It may be noted that the product estimator $\bar{y}_p$, defined in (7.34), can be corrected for bias. The resulting estimator

$$\bar{y}_p' = \frac{\bar{y}\,\bar{x}}{\bar{X}} - \left(\frac{N-n}{Nn}\right)\left(\frac{s_{xy}}{\bar{X}}\right)$$

is unbiased for the population mean $\bar{Y}$. Some other unbiased product type strategies have also been developed by Gupta and Adhvaryu (1982). Also, some of the procedures mentioned in remarks 7.3 and 7.4 of section 7.6 can be used to obtain unbiased product type estimators.

Example 7.8

A psychologist needs information on average duration of sleep (in hours) during night for the persons equal, or over, 50 years of age in a certain locality. Voters' list for the

locality is used to prepare the list of persons aged 50 years or more. This population frame has 546 such persons. A WOR equal probability sample of 30 persons is drawn. Information regarding the age and duration of sleep gathered from the sampled persons is given in table 7.5.

Table 7.5 Age and average duration of sleep (in hours) for sample persons

Person	Age (x)	Duration of sleep (y)	Person	Age (x)	Duration of sleep (y)
1	62	7.00	16	66	7.00
2	75	5.00	17	78	5.75
3	51	8.00	18	63	6.75
4	57	7.75	19	77	5.50
5	81	5.00	20	73	4.75
6	79	5.25	21	55	7.30
7	67	7.00	22	71	6.00
8	74	6.25	23	63	6.50
9	84	4.50	24	87	4.50
10	56	7.75	25	61	6.25
11	68	7.00	26	58	6.25
12	70	7.00	27	60	6.50
13	59	7.25	28	69	6.00
14	64	6.75	29	56	6.50
15	53	8.50	30	71	5.75

The average age of persons in the target population is 70 years. Treating age as the auxiliary variable, estimate average duration of sleep in the population. Build up the confidence interval for it, and also, estimate percent relative efficiency of this estimator in relation to the SRS based usual mean estimator $\bar{y}$.

Solution

We have N= 546 and n=30. The average duration of sleep for respondents is likely to decrease with increase in age. In order to be sure that the product method of estimation could work well, we first work out

$$\bar{y} = 6.38,\ \bar{x} = 66.93,\ \bar{X} = 70,\ \hat{R} = \bar{y}/\bar{x} = .09532,$$

$$s_y^2 = 1.084,\ s_x^2 = 92.409, \text{ and } s_{xy} = -8.885.$$

These figures in turn yield the sample estimate of the correlation coefficient as

$$r = \frac{s_{xy}}{\sqrt{s_y^2\, s_x^2}}$$

$$= \frac{-8.885}{\sqrt{(1.084)\,(92.409)}}$$

$$= -.89$$

In the light of condition (7.42), the product method of estimation could be used profitably. Thus, we get from (7.34), the estimate of average duration of sleep as

$$\bar{y}_p = \frac{\bar{y}\,\bar{x}}{\bar{X}}$$

$$= \frac{(6.38)\,(66.93)}{70}$$

$$= 6.10$$

For building up the confidence interval, we require estimate of mean square error. It is computed by using (7.37). Therefore,

$$\text{mse}\,(\bar{y}_p) = \left(\frac{1}{n} - \frac{1}{N}\right)(s_y^2 + \hat{R}^2 s_x^2 + 2\hat{R}\, s_{xy})$$

$$= \left(\frac{1}{30} - \frac{1}{546}\right)[1.084 + (.09532)^2\,(92.409) + 2(.09532)\,(-8.885)]$$

$$= .00724$$

The confidence interval for average duration of sleep in the population is given by

$$\bar{y}_p \pm 2\sqrt{\text{mse}\,(\bar{y}_p)}$$

$$= 6.10 \pm 2\sqrt{.00724}$$

$$= 5.93,\ 6.27$$

Thus, the persons of age 50 years and more are, on the average, likely to sleep from 5.93 to 6.27 hours.

The estimated percent relative efficiency of the product estimator $\bar{y}_p$, in relation to the simple mean estimator $\bar{y}$, is given by

$$\text{RE} = \frac{v(\bar{y})}{\text{mse}\,(\bar{y}_p)}\,(100)$$

where from (3.10)

$$v(\bar{y}) = \frac{N-n}{Nn}\,s_y^2$$

$$= \frac{(546-30)}{(546)\,(30)}\,(1.084)$$

$$= .03415$$

Hence,

$$RE = \frac{.03415}{.00724}(100)$$

$$= 471.69$$

Therefore, the use of auxiliary information on age, in the form of product estimator, has resulted in increasing the relative efficiency from 100% to 471.69%. ■

7.8 DETERMINATION OF SAMPLE SIZE FOR PRODUCT ESTIMATOR

In order to arrive at the required sample size for estimating mean/total, let n_1 be the number of units selected in the preliminary sample. The value of the expression $(s_y^2 + \hat{R}^2 s_x^2 + 2\hat{R} s_{xy})$ in (7.37), is calculated from the observations on this preliminary sample. Let this value be denoted by

$$s_p^2 = (s_{y1}^2 + \hat{R}_1^2 s_{x1}^2 + 2\hat{R}_1 s_{xy1}) \tag{7.43}$$

The mean square error estimator then becomes $mse(\bar{y}_p) = [(N-n)/Nn]\, s_p^2$, and $mse(\hat{Y}_p) = [N(N-n)/n]\, s_p^2$. Using these calculated estimates in place of $mse(\bar{y}_p)$ and $mse(\hat{Y}_p)$ in (7.37) and (7.41) respectively, and proceeding in the same way as in case of ratio estimator of mean/total, one gets the rule for choosing the required sample size for estimating the population mean/total through product method of estimation.

Sample size to estimate the population mean/total with a bound B on the error of estimation :

$$n = \frac{Ns_p^2}{ND + s_p^2} \tag{7.44}$$

where

$$D = \frac{B^2}{4} \quad \text{(when estimating mean)}$$

$$D = \frac{B^2}{4N^2} \quad \text{(when estimating total)}$$

and s_p^2 is as defined in (7.43). When $n_1 \geq n$, no additional unit need be selected, otherwise, augment the preliminary sample by selecting $(n-n_1)$ more units.

Example 7.9

Assume that the sample of 30 persons, drawn in example 7.8, is a preliminary sample. Examine, whether this sample is sufficient for estimating mean sleeping hours with a margin of error .25 hours ?

Solution

We have assumed the sample of 30 persons drawn in example 7.8 as the preliminary sample. Obviously, all the estimated values computed in that example will be treated as estimates provided by the preliminary sample. We, therefore, get

$$N = 546,\ n_1 = 30,\ \hat{R} = \hat{R}_1 = .09532,\ s_y^2 = s_{y1}^2 = 1.084,$$

$$s_x^2 = s_{x1}^2 = 92.409, \text{ and } s_{xy} = s_{xy1} = -8.885.$$

Using (7.43), we first work out s_p^2. Thus, we have

$$\begin{aligned} s_p^2 &= s_{y1}^2 + \hat{R}_1^2\, s_{x1}^2 + 2\hat{R}_1\, s_{xy1} \\ &= 1.084 + (.09532)^2\,(92.409) + 2\,(.09532)\,(-8.885) \\ &= .2298 \end{aligned}$$

Now,

$$D = \frac{B^2}{4} = \frac{(.25)^2}{4} = .015625$$

The required sample size is then obtained by using (7.44). We thus have

$$\begin{aligned} n &= \frac{Ns_p^2}{ND + s_p^2} \\ &= \frac{(546)\,(.2298)}{(546)\,(.015625) + .2298} \\ &= 14.3 \\ &\approx 14 \end{aligned}$$

As the preliminary sample size $n_1 = 30$ is more than the required sample size $n = 14$, the sample of size 30 already drawn is sufficient to yield an estimate of mean sleeping hours with desired accuracy. ■

The reader should note that like ratio estimator, the product estimator can also be used in stratified simple random sampling yielding separate and combined product estimators. Thus we have

$$\bar{y}_{sp} = \sum_{h=1}^{L} W_h \left(\frac{\bar{y}_h \bar{x}_h}{\bar{X}_h} \right) \tag{7.45}$$

and

$$\bar{y}_{cp} = \frac{1}{\bar{X}} \left(\sum_{h=1}^{L} W_h \bar{y}_h \right) \left(\sum_{h=1}^{L} W_h \bar{x}_h \right) \tag{7.46}$$

Both the above estimators are biased and the expressions for their biases are given by

$$B(\bar{y}_{sp}) = \sum_{h=1}^{L} W_h \left(\frac{N_h - n_h}{N_h n_h} \right) \frac{S_{hxy}}{\bar{X}_h} \tag{7.47}$$

and

$$B(\bar{y}_{cp}) = \frac{1}{\bar{\bar{X}}} \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h} \right) S_{hxy} \tag{7.48}$$

Expressions for mean square errors and their estimators for the above two estimators can be obtained from the corresponding expressions for separate and combined ratio estimators by replacing $(-2R_h S_{hxy})$, $(-2\hat{R}_h s_{hxy})$, $(-2RS_{hxy})$, and $(-2\hat{R}\, s_{hxy})$ in (7.24), (7.25), (7.28), and (7.29) respectively with same terms with a positive sign.

LET US DO

7.1 It is desired to estimate per acre yield of wheat crop, which is the ratio of total wheat yield to the total area under the crop, in a certain area. Discuss, how will you go about it ?

7.2 An investigator wishes to estimate sex ratio in a town. The best frame available is the list of ration depots (shops from where people get essential commodities at prices fixed by the government). Each household possesses a ration card, and is listed with a nearby depot. The details of family members, like sex, age, etc., are mentioned on the ration card. These details are also available with the ration depot. Twenty seven depots were selected from a total of 148 depots in the town, using SRS without replacement. The information in respect of family members gathered from the selected depots is given below, where M and F stand for number of males and females respectively.

Depot	M	F	Depot	M	F	Depot	M	F
1	870	630	10	911	860	19	890	763
2	500	440	11	731	601	20	525	624
3	981	670	12	508	520	21	560	574
4	893	703	13	680	570	22	674	766
5	613	688	14	713	591	23	990	720
6	380	360	15	507	569	24	863	618
7	421	473	16	984	887	25	782	614
8	671	576	17	768	703	26	828	701
9	933	956	18	635	648	27	774	540

Estimate the male : female ratio in the town, and work out confidence interval for it.

7.3 A survey was undertaken to estimate change in the per acre rental value of irrigated land in a certain district comprising of 760 villages. A WOR simple random sample of 30 villages was drawn. *Sarpanch*, the elected head of each village, was interviewed and the information on rental value of irrigated land for the current year (y) and the assessed rental value 5 years back (x) was obtained. The collected information (in '00 rupees) is given below :

Village	x	y	Village	x	y	Village	x	y
1	35	48	11	35	46	21	33	48
2	40	46	12	37	49	22	34	46
3	38	50	13	34	45	23	37	49
4	42	51	14	38	49	24	34	47
5	38	47	15	35	44	25	33	46
6	37	46	16	34	46	26	35	46
7	35	41	17	33	45	27	34	47
8	36	43	18	37	49	28	33	45
9	35	44	19	34	48	29	34	49
10	36	47	20	36	42	30	36	50

Using the sample information, estimate the change in rental value, and place confidence limits on it.

7.4 Describe ratio method for estimating population total, and state the situations where the estimator based on this method is likely to be more efficient than the usual SRS estimator. Under what condition the bias of this estimator becomes zero ?

7.5 A cattle-cake manufacturer is interested in examining the effect of a new cattle-cake on milk yield. For this purpose, a WOR simple random sample of 27 buffalos was drawn from a population of 540 buffalos. One day milk yield of these 27 buffalos was recorded individually, and the milk yield for the remaining 540-27= 513 buffalos was measured collectively. This gave the average per day milk yield for the population of 540 buffalos as 10 kg. The selected 27 buffalos were then kept on the new feed for 10 days. After the expiry of this period, the milk yield for a day of these sample buffalos was again recorded. The average milk yield (in kg) for the sample buffalos was then computed. The milk yields denoted by x and y respectively, recorded before and after the introduction of the new feed, for the selected buffalos are given as follows :

Buffalo	Milk yield		Buffalo	Milk yield		Buffalo	Milk yield	
	x	y		x	y		x	y
1	10.8	11.6	10	12.4	13.1	19	8.8	9.6
2	8.6	9.4	11	8.9	10.2	20	6.5	7.9
3	7.5	8.3	12	11.3	12.0	21	14.1	14.5
4	11.4	12.9	13	12.0	12.8	22	9.1	10.4
5	12.9	13.4	14	9.6	10.9	23	13.6	14.2
6	7.7	8.5	15	7.5	8.7	24	9.4	10.1
7	9.6	10.9	16	13.6	14.3	25	10.7	10.9
8	10.3	11.0	17	8.5	9.4	26	8.4	9.6
9	11.6	12.2	18	14.5	14.8	27	11.8	12.2

Estimate average daily milk yield per buffalo after being kept on the new feed, and place confidence limits on it. Also, work out the estimated relative efficiency of the ratio estimator of average daily milk yield with respect to the usual estimator $\bar{y}$ based on SRS without replacement.

7.6 A town consists of 138 wards. Total number of dwellings in the town are known to be 14060. An investigator desires to estimate the total number of dwellings occupied by tenants. For this purpose, a WOR simple random sample of 24 wards was selected. The information on number of dwellings (x), and the number of dwellings occupied by tenants (y), was obtained for the selected wards. This is presented in table below :

Ward	x	y	Ward	x	y	Ward	x	y
1	80	8	9	116	9	17	93	8
2	218	16	10	130	11	18	157	13
3	108	4	11	105	12	19	87	6
4	90	6	12	75	6	20	137	10
5	126	10	13	128	10	21	77	7
6	143	16	14	110	9	22	179	21
7	70	6	15	85	9	23	198	16
8	85	5	16	150	13	24	166	13

Using ratio method, estimate the total number of rented dwellings in the town, and determine confidence limits for it. Also, estimate relative efficiency of the ratio estimator of this total in relation to the usual SRS without replacement based estimator $\hat{Y}$.

7.7 Given a preliminary sample of size n_1, discuss the procedure of determining the necessary sample size for estimating population ratio R, when the margin of error that can be tolerated is 4%.

7.8 Assume that the sample of 27 depots, drawn in exercise 7.2, is a preliminary sample. Using the information presented in that exercise, verify whether this sample size is sufficient to estimate male : female ratio with a margin of error .05 ? If not, how many additional units need to be selected?

7.9 Treating the estimate of mean square error obtained in exercise 7.6 as from a preliminary sample, check whether the sample of 24 wards is sufficient to estimate the total number of dwellings occupied by tenants with a margin of error as 30 ?

7.10 The students admitted to a college this year were stratified into 3 categories - rich, middle class, and poor - depending on the income and occupation of their parents, they had stated in their admission forms. The three strata respectively consisted of $N_1 = 120$, $N_2 = 300$, and $N_3 = 390$ students. The average annual income of the parents of students belonging to rich, middle class, and poor strata was worked out from the admission forms of the students. It was found to be rupees 92, 60, and 40 thousands respectively. The investigator wishes to estimate average amount of pocket money these students had spent during the preceding 3 months. It was decided to select a WOR simple random sample of size 27 from the total population of 810 students. Using proportional allocation,

$$n_1 = \left(\frac{27}{810}\right)(120) = 4$$

$$n_2 = \left(\frac{27}{810}\right)(300) = 10$$

$$n_3 = \left(\frac{27}{810}\right)(390) = 13$$

students were selected from the first, second, and third stratum respectively. The information regarding pocket money (y) they had spent during the past three months was obtained through personal interview. Given below is the amount of pocket money spent (in rupees) and the annual income (x) of students' parents (in '000 rupees).

Rich		Middle class				Poor			
x	y	x	y	x	y	x	y	x	y
78	900	50	450	62	550	38	400	43	200
135	1250	66	700	50	450	41	350	33	250
70	750	58	650	56	500	47	450	30	200
87	800	60	500	63	650	43	400	46	400
		55	550	52	350	34	250	37	350
						35	200	39	350
						28	150		

Using combined ratio estimator, estimate the average pocket money a student had spent during the past three months, and build up the confidence interval for it.

7.11 From the data presented in exercise 7.10, estimate the average amount of pocket money spent by a student using separate ratio estimator. Also, discuss which of the two ratio estimators - separate or combined - will be more appropriate in the present situation ?

7.12 In which situation the product estimator of mean is more efficient than the simple random sample mean ? Do you agree with the statement that the product method of estimation finds application in lesser number of situations commonly encountered in practice in relation to the ratio method ? If so, why ?

7.13 Writing the y values in example 7.2 in decreasing order while retaining x values as such, compare the mean square error of product estimator $\bar{y}_p$ with the variance of the usual SRS mean estimator $\bar{y}$ based on WOR sample of size 3. Also, work out percent relative efficiency of the estimator $\bar{y}_p$ with respect to $\bar{y}$.

7.14 It is of interest to estimate weekly time spent by the undergraduate students of a university in viewing television, playing cards, gossips, or just wandering around, etc. In order to accomplish this task, a WOR simple random sample of 36 students was drawn from a population of 960 students. The overall grade point average (OGPA) was used as auxiliary variable. The mean OGPA for this population, obtained from the office of Registrar, is 2.97 (4.00 basis). The information collected in respect of OGPA (x), and the weekly time (in hours) spent on the above said nonacademic activities (y), is presented in the table below :

Student	x	y	Student	x	y	Student	x	y
1	2.35	28	13	1.91	20	25	3.75	6
2	2.00	33	14	2.84	22	26	3.35	8
3	2.75	14	15	2.36	17	27	2.95	12
4	2.80	21	16	3.62	10	28	3.09	11
5	2.95	17	17	1.83	18	29	2.81	12
6	2.70	25	18	2.69	14	30	2.71	13
7	3.43	10	19	3.29	11	31	2.92	13
8	3.82	7	20	2.41	19	32	3.36	9
9	3.56	16	21	2.07	13	33	1.96	20
10	2.55	15	22	1.86	18	34	3.66	8
11	2.85	14	23	2.79	13	35	3.23	7
12	3.44	7	24	2.71	11	36	1.81	20

Using appropriate method, estimate average weekly time (in hours) spent by a student on nonacademic activities. Also, build up the confidence interval for this average.

7.15 Using data of exercise 7.14, estimate total weekly time (in hours) spent in nonacademic activities by all the students in the population under study. Also, build up the confidence interval for this total time spent.

7.16 Using estimate of mean square error obtained in exercise 7.14, comment whether, or not, the sample of 36 students is sufficient to provide an estimate of average weekly time spent by a student on nonacademic activities with a margin of error equal to 2 hours. If not, how many more students should be included in the sample ?

CHAPTER 8

Regression Method of Estimation

8.1 INTRODUCTION

Analogous to the ratio and product estimators, the linear *regression estimator* is also designed to increase the efficiency of estimation by using information on the auxiliary variable x which is correlated with the study variable y. As stated before, the ratio method of estimation is at its best when the correlation between y and x is positive and high, and also the regression of y on x is linear through the origin. In practice, however, it is observed that even when the regression of y on x is linear, the regression line passes through a point away from the origin. The efficiency of the ratio estimator in such cases is very low, as it decreases with the increase in length of the intercept cut on y-axis by the regression line. Regression estimator is the appropriate estimator for such situations. Although this estimator requires little more calculations than the ratio estimator, it is always at least as efficient as the ratio estimator for estimating population mean or total. Similarly, the product estimator of population mean or total is never more efficient than the corresponding linear regression estimator.

Regression estimator for mean/total has been in use for quite sometime. Watson (1937) used regression of leaf area on leaf weight to estimate the average leaf area for a plant. In another interesting application presented by Yates (1960), an eye estimate of the volume of timber was made on each of the 1/10 acre plots of a population of plots. The actual timber volume was measured for a sample of plots. Using the regression estimator, total timber volume was then estimated.

In this chapter, we shall discuss two types of regression estimators - one when the value of the regression coefficient is known in advance, and the other when it is to be estimated from the sample. The estimator based on the known value of regression coefficient is termed difference estimator.

8.2 ESTIMATION OF MEAN/TOTAL USING DIFFERENCE ESTIMATOR

We assume that (y_1, x_1), (y_2, x_2), ...,(y_n, x_n) denote the observations on the study and auxiliary variables on an SRS without replacement sample of n units selected from the finite population of N units. Also, let $\bar{y}$ and $\bar{x}$ denote the corresponding sample means, whereas $\bar{X}$ is the population mean for the auxiliary variable. Let α be a predetermined constant. Then, the *difference estimator* of population mean $\bar{Y}$ is defined as

$$\bar{y}_d = \bar{y} + \alpha(\bar{X} - \bar{x})$$

The estimator $\bar{y}_d$ is unbiased for $\bar{Y}$, whatever be the value of α. It is easy to see that the variance $V(\bar{y}_d)$ is minimum when

$$\alpha = \frac{S_{xy}}{S_x^2} = \beta \text{ (say)}$$

where β is the population regression coefficient of y on x. In practice, the actual value of β will not be available. Efficiency considerations will, however, require that a value of α very close to β be chosen for building up the difference estimator.

In repeated surveys, analysis of survey data over time may indicate that the population regression coefficient remains fairly constant. It may help us in choosing an appropriate value for the regression coefficient of y on x, in advance. The difference estimators, based on predetermined value β_o of the regression coefficient β, are simple and informative. We first consider these estimators. The subscript (d) used with the estimator stands for the difference estimator.

Unbiased difference estimator of population mean $\bar{Y}$:

$$\bar{y}_d = \bar{y} + \beta_o (\bar{X} - \bar{x}) \tag{8.1}$$

Variance of estimator $\bar{y}_d$:

$$V(\bar{y}_d) = \frac{N-n}{Nn} (S_y^2 + \beta_o^2 S_x^2 - 2\beta_o S_{xy})$$

$$= \frac{N-n}{Nn} S_y^2 (1-\rho^2) \quad \text{when } \beta_o = \beta \tag{8.2}$$

Estimator of variance $V(\bar{y}_d)$:

$$v(\bar{y}_d) = \frac{N-n}{Nn} (s_y^2 + \beta_o^2 s_x^2 - 2\beta_o s_{xy}) \tag{8.3}$$

The parameters and their respective estimators involved in (8.1) to (8.3), have already been defined in the preceding chapter.

From $V(\bar{y}_d)$ in (8.2) and $V(\bar{y})$ in (3.9), it is obvious that regression estimator using known β_o is always more efficient than the usual simple random sample mean when $\beta_o = \beta$. The efficiency of the estimator $\bar{y}_d$ in (8.1) will decrease, as the difference between the guess value β_o and the actual value β of the regression coefficient increases.

As stated earlier, the corresponding expressions for estimating population total can be easily obtained.

Example 8.1

A medical student was given an assignment to estimate the average systolic blood pressure (BP) for teachers between 30 and 60 years of age in a certain university. The objective was to compare it with the systolic BP of a part of the population engaged in manual work. A WOR simple random sample of 24 teachers was drawn from the frame consisting of 961 teachers. Age of teachers was taken as the auxiliary variable. Average

age for the population of teachers was calculated from university records as 42.7 years. The regression coefficient of systolic BP on age, available from an earlier study conducted 5 years ago, is .8952. Using difference estimator, estimate the average systolic blood pressure, and build up confidence interval for it.

Table 8.1 Age (in completed years) and systolic blood pressure for sample teachers

Teacher	BP (y)	Age (x)	Teacher	BP (y)	Age (x)	Teacher	BP (y)	Age (x)
1	130	38	9	130	58	17	146	55
2	128	41	10	124	36	18	144	49
3	147	55	11	150	57	19	126	35
4	130	37	12	125	31	20	124	32
5	128	39	13	121	36	21	148	58
6	120	30	14	115	47	22	140	34
7	151	48	15	163	59	23	133	43
8	140	44	16	141	47	24	143	50

Solution

Here, we have $N = 961$, $n = 24$, $\bar{X} = 42.7$, and $\beta_o = .8952$. We first work out intermediate sample values to be used later. These are

$$\bar{y} = \frac{1}{24}(130 + 128 + \ldots + 143)$$

$$= 135.29$$

$$\bar{x} = \frac{1}{24}(38 + 41 + \ldots + 50)$$

$$= 44.13$$

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

$$= \frac{1}{23}[(130)^2 + (128)^2 + \ldots + (143)^2 - 24(135.29)^2]$$

$$= 146.08$$

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

$$= \frac{1}{23}[(38)^2 + (41)^2 + \ldots + (50)^2 - 24(44.13)^2]$$

$$= 89.13$$

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

$$= \frac{1}{23}\left[(38)(130) + (41)(128) + \ldots + (50)(143) - 24(44.13)(135.29)\right]$$

$$= \frac{1}{23}(145194 - 143288.34)$$

$$= 82.85$$

For working out estimate of average systolic BP, we use (8.1). That means,

$$\bar{y}_d = \bar{y} + \beta_o(\bar{X} - \bar{x})$$

$$= 135.29 + .8952(42.7 - 44.13)$$

$$= 134.01$$

Next step is to estimate the variance of the estimator $\bar{y}_d$. For this, we use (8.3). Thus,

$$v(\bar{y}_d) = \frac{N-n}{Nn}(s_y^2 + \beta_0^2 s_x^2 - 2\beta_o s_{xy})$$

$$= \frac{961-24}{(961)(24)}\left[146.08 + (.8952)^2(89.13) - 2(.8952)(82.85)\right]$$

$$= 2.8102$$

The confidence limits, in which the average systolic BP is likely to fall, are

$$\bar{y}_d \pm 2\sqrt{v(\bar{y}_d)}$$

$$= 134.01 \pm 2\sqrt{2.8102}$$

$$= 134.01 \pm 3.35$$

$$= 130.66,\ 137.36$$

Hence, the average systolic blood pressure for the population of 961 teachers is most likely to take a value in the closed interval [130.66, 137.36].

If the information on the auxiliary variable x (age) was not used, the average systolic blood pressure could be estimated through the usual simple random sample mean $\bar{y}$. The variance of this estimator for WOR sampling would have been estimated from (3.10) as

$$v(\bar{y}) = \frac{N-n}{Nn} s_y^2$$

$$= \frac{961-24}{(961)(24)}(146.08)$$

$$= 5.9347$$

which is larger than $v(\bar{y}_d)$. The percent relative efficiency of $\bar{y}_d$ with respect to $\bar{y}$, is estimated by

$$RE = \frac{v(\bar{y})}{v(\bar{y}_d)}(100)$$

$$= \frac{5.9347}{2.8102}(100)$$

$$= 211.18$$

Thus, the use of auxiliary information on age, in the form of difference estimator, has increased the efficiency more than two times for the estimation of average systolic blood pressure. ■

8.3 ESTIMATION OF MEAN/TOTAL USING ESTIMATED REGRESSION COEFFICIENT

The estimator $\bar{y}_d$ considered in the preceding section is unbiased for $\bar{Y}$ and the expression for its variance is exact. The estimator of variance is also unbiased. The estimator itself doesn't require heavy computations and is easy to apply in practice. Its relative precision depends on the accuracy with which the value of the regression coefficient β has been guessed. However, situations are frequently encountered where it is not possible to make a reliable guess for β. Then, the only alternative left is to estimate it from the sample itself and use the estimated value in place of β_o, in (8.1). The estimator of population mean $\bar{Y}$, so obtained, is called *linear regression estimator.*

From least square principle, the sample estimate for β is given by

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}$$

Since this estimator $\hat{\beta}$ is a random variable, exact expressions for the expected value and the mean square error of the regression estimator are bit difficult to obtain. We restrict again to the WOR simple random sampling and assume that the sample is large enough, so that, the approximate expressions for $MSE(\bar{y}_{lr})$ and its estimator in (8.6) and (8.7) are accurate enough.

Regression estimator of population mean $\bar{Y}$:

$$\bar{y}_{lr} = \bar{y} + \hat{\beta}(\bar{X} - \bar{x}) \tag{8.4}$$

Bias of estimator $\bar{y}_{lr}$:

$$B(\bar{y}_{lr}) = -\operatorname{Cov}(\bar{x}, \hat{\beta}) \tag{8.5}$$

Approximate mean square error of estimator $\bar{y}_{lr}$:

$$MSE(\bar{y}_{lr}) = \frac{N-n}{Nn}(S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy})$$

$$= \frac{N-n}{Nn} S_y^2(1-\rho^2) \tag{8.6}$$

Estimator of mean square error MSE ($\bar{y}_{lr}$) :

$$\text{mse}(\bar{y}_{lr}) = \frac{N-n}{Nn}\left(s_y^2 + \hat{\beta}^2 s_x^2 - 2\hat{\beta} s_{xy}\right)$$

$$= \frac{N-n}{Nn} s_y^2 (1-r^2) \qquad (8.7)$$

where r is the sample correlation coefficient.

The corresponding expressions for the estimator of population total can be obtained from the above expressions in the usual manner.

Example 8.2

A physiologist has undertaken a project to estimate the average leaf area of a newly developed strain of wheat of which 120 plants were raised. In all, there were 2106 leaves and their total weight was 242.078 gm. Since measuring of area of all the 2106 leaves is difficult, a WOR simple random sample of 33 leaves was drawn. The area and weight of each sampled leaf were recorded. These are given in table 8.2 below. Estimate the average leaf area for the population under consideration. Also, work out the confidence interval for the actual value of this population parameter.

Table 8.2 Area (in sq cm) and weight (in mg) of leaves in the sample

Leaf	Area (y)	Weight (x)	Leaf	Area (y)	Weight (x)	Leaf	Area (y)	Weight (x)
1	27.37	105	12	24.18	106	23	14.31	78
2	30.21	109	13	35.72	125	24	39.28	128
3	22.18	100	14	19.76	97	25	24.16	102
4	36.76	125	15	33.46	125	26	26.51	114
5	28.51	116	16	43.62	131	27	29.69	119
6	30.34	118	17	16.11	85	28	20.03	101
7	21.81	104	18	21.07	112	29	18.41	93
8	29.11	118	19	26.71	117	30	35.72	124
9	38.90	129	20	18.51	96	31	29.33	117
10	17.21	90	21	23.43	103	32	21.88	107
11	42.44	130	22	31.66	121	33	21.29	103

Solution

We have N = 2106, n = 33, and X = 242.078 gm. This implies that

$$\bar{X} = \frac{242.078}{2106}$$

$$= .1149 \text{ gm}$$

$$= 114.9 \text{ mg}$$

As in examples 7.2, 7.3, and 8.1, we first work out the intermediate sample estimates. These are

$\bar{y} = 27.263$, $\bar{x} = 110.545$, $s_y^2 = 61.003$, $s^2{}_x = 187.631$, and $s_{xy} = 100.312$

Thus,

$$r = \frac{s_{xy}}{s_y s_x}$$

$$= \frac{100.312}{\sqrt{(61.003)(187.631)}}$$

$$= .9376$$

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}$$

$$= \frac{100.312}{187.631} = .5346$$

Now, we compute the estimate of average leaf area by using (8.4). Hence,

$$\bar{y}_{lr} = \bar{y} + \hat{\beta}(\bar{X} - \bar{x})$$

$$= 27.263 + .5346(114.9 - 110.545)$$

$$= 29.591$$

Estimate of mean square error is calculated from (8.7), as

$$mse(\bar{y}_{lr}) = \frac{N-n}{Nn}(s_y^2 + \hat{\beta}^2 s_x^2 - 2\hat{\beta} s_{xy})$$

$$= \frac{2106-33}{(2106)(33)}[61.003 + (.5346)^2 (187.631) - 2(.5346)(100.312)]$$

$$= .2200$$

Alternatively, it can also be computed by using second version of (8.7). This yields

$$mse(\bar{y}_{lr}) = \frac{N-n}{Nn} s_y^2 (1 - r^2)$$

$$= \frac{2106-33}{(2106)(33)}(61.003)[1 - (.9376)^2]$$

$$= .2200$$

The range, in which the average leaf area for the population would probably lie, is determined by using confidence limits. These are obtained as

$$\bar{y}_{lr} \pm 2\sqrt{mse(\bar{y}_{lr})}$$

$$= 29.591 \pm 2\sqrt{.2200}$$

$$= 28.65, 30.53$$

Thus, the average leaf area is expected to be in the range 28.65 to 30.53 sq cm, with probability approximately equal to .95.

It should be noted here that if the information on the auxiliary variable x was not used in the form of the regression estimator and the average leaf area was estimated using the simple sample mean $\bar{y}$, the estimate of the variance $V(\bar{y})$ from (3.10) would have been

$$v(\bar{y}) = \frac{N-n}{Nn} s_y^2$$

$$= \frac{2106-33}{(2106)(33)} (61.003)$$

$$= 1.8196$$

which is much larger than $mse(\bar{y}_{lr})$. The estimated percent relative efficiency of the linear regression estimator, with respect to the simple mean $\bar{y}$, is given by

$$RE = \frac{v(\bar{y})}{mse(\bar{y}_{lr})} (100)$$

$$= \frac{1.8196}{.2200} (100)$$

$$= 827.09$$

Hence, we find that the linear regression estimator is over 8 times more efficient than the simple mean $\bar{y}$. One could also say, that the use of auxiliary information, in the form of regression estimator, has reduced the error in the estimate of average leaf area by over eight times. ■

8.4 SAMPLE SIZE DETERMINATION FOR ESTIMATING MEAN / TOTAL

For estimating the size of the sample required to estimate population mean/total with a specified bound B on the error of estimation, we proceed as in chapter 7. Let a preliminary sample of n_1 units be selected using SRS without replacement. The values of $(s_y^2 + \beta_o^2 s_x^2 - 2\beta_o s_{xy})$ and $(s_y^2 + \hat{\beta}^2 s_x^2 - 2\hat{\beta} s_{xy})$ are obtained using data from this preliminary sample, and are redenoted as

$$s_{11}^2 = s_{y1}^2 + \beta_o^2 s_{x1}^2 - 2\beta_o s_{xy1}$$

$$s_{12}^2 = s_{y1}^2 + \hat{\beta}_1^2 s_{x1}^2 - 2\hat{\beta}_1 s_{xy1}$$

respectively. These values are then used in $v(\bar{y}_d)$ and $mse(\bar{y}_{lr})$ given in (8.3) and (8.7) respectively, in place of the corresponding values that would have been obtained from the sample of n units. The equations

$$2\sqrt{v(\bar{y}_d)} = B \qquad \text{(for difference estimator)} \qquad (8.8)$$

$$2\sqrt{mse(\bar{y}_{lr})} = B \qquad \text{(for linear regression estimator)} \qquad (8.9)$$

are then solved for n.

For estimating total, $v(\bar{y}_d)$ and $mse(\bar{y}_{lr})$ in the above two equalities are multiplied by N^2. Depending on the estimator used, the solution of (8.8) or (8.9) for n, gives the formula for determining the required sample size when estimating population mean/total. Thus we have :

Sample size required to estimate the population mean/total with a bound B on the error of estimation :

$$n = \frac{Ns_{li}^2}{ND + s_{li}^2}, \quad i = 1, 2 \qquad (8.10)$$

where

$$D = \frac{B^2}{4} \quad \text{(when estimating mean)}$$

$$D = \frac{B^2}{4N^2} \quad \text{(when estimating total)}$$

$$s_{l1}^2 = s_{y1}^2 + \beta_o^2 s_{x1}^2 - 2\beta_o s_{xy1} \quad \text{(in case of difference estimator)}$$

$$s_{l2}^2 = s_{y1}^2 + \hat{\beta}_1^2 s_{x1}^2 - 2\hat{\beta}_1 s_{xy1} \quad \text{(for linear regression estimator)}$$

If $n_1 \geq n$, the sample size n_1 is sufficient. Otherwise, $(n-n_1)$ more units will have to be selected in the sample.

Example 8.3

Assume that the sample drawn in example 8.2 is a preliminary sample. Based on information from this sample, determine the sample size required to estimate the average leaf area with bound on the error of estimation as one sq cm.

Solution

Here, we have N=2106 and B=1. The sample of 33 plants selected in example 8.2 is now taken as the preliminary sample. All the sample estimates computed in example 8.2 will, therefore, be treated as preliminary sample estimates. Thus,

$$n_1 = 33, \ \hat{\beta} = \hat{\beta}_1 = .5346, \ s_y^2 = s_{y1}^2 = 61.003,$$

$$s_x^2 = s_{x1}^2 = 187.631, \text{ and } s_{xy} = s_{xy1} = 100.312.$$

Then we compute

$$\begin{aligned} s_{l2}^2 &= s_{y1}^2 + \hat{\beta}_1^2 s_{x1}^2 - 2\hat{\beta}_1 s_{xy1} \\ &= 61.003 + (.5346)^2 (187.631) - 2(.5346)(100.312) \\ &= 7.374 \end{aligned}$$

Further,

$$D = \frac{B^2}{4} = \frac{1}{4} = .25$$

The required sample size can then be obtained from (8.10). This means,

$$n = \frac{Ns_{12}^2}{ND + s_{12}^2}$$

$$= \frac{(2106)(7.374)}{(2106)(.25) + 7.374}$$

$$= 29.1$$

$$\approx 29$$

As $n_1 > n$, the already selected sample of 33 leaves is sufficient to achieve the required margin of error. ■

8.5 SEPARATE AND COMBINED REGRESSION ESTIMATORS

As with the ratio estimator, two types of regression estimators can be developed for stratified random sampling. In the first estimator $\bar{y}_{lrs}$, a *separate regression estimate* is computed for each stratum mean and then their weighted average is taken. This estimator is appropriate, when one suspects that the true regression coefficients β_h, h=1,2,..., L, vary from stratum to stratum. As in chapters 5 and 7, let n_h be the number of units selected in the sample from the h-th stratum, containing N_h units, using SRS without replacement. Let

$$\rho_h = \frac{S_{hxy}}{S_{hy} S_{hx}}$$

$$r_h = \frac{s_{hxy}}{s_{hy} s_{hx}}$$

denote the correlation coefficients for the h-th stratum computed from N_h and n_h units respectively. Further, let

$$\beta_h = \frac{S_{hxy}}{S_{hx}^2}$$

be the unknown regression coefficient of y on x in the h-th stratum, and

$$\hat{\beta}_h = \frac{s_{hxy}}{s_{hx}^2} \tag{8.11}$$

its least square estimator from the sample of size n_h, h = 1, 2,..., L. We shall also assume that the sample sizes $\{n_h\}$ are large enough, so that, the approximations used in $B(\bar{y}_{lrs})$, $MSE(\bar{y}_{lrs})$, and estimator $mse(\bar{y}_{lrs})$ are sufficiently accurate.

Separate regression estimator of population mean $\bar{Y}$:

$$\left.\begin{aligned} \bar{y}_{lrs} &= \sum_{h=1}^{L} W_h \bar{y}_{hlr} \\ &= \sum_{h=1}^{L} W_h [\bar{y}_h + \hat{\beta}_h (\bar{X}_h - \bar{x}_h)] \end{aligned}\right] \quad (8.12)$$

Bias of estimator $\bar{y}_{lrs}$:

$$B(\bar{y}_{lrs}) = - \sum_{h=1}^{L} W_h \, Cov(\bar{x}_h, \hat{\beta}_h) \quad (8.13)$$

Approximate mean square error of estimator $\bar{y}_{lrs}$:

$$\left.\begin{aligned} MSE(\bar{y}_{lrs}) &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right)(S_{hy}^2 + \beta_h^2 S_{hx}^2 - 2\beta_h S_{hxy}) \\ &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) S_{hy}^2 (1 - \rho_h^2) \end{aligned}\right] \quad (8.14)$$

Estimator of mean square error MSE ($\bar{y}_{lrs}$) :

$$\left.\begin{aligned} mse(\bar{y}_{lrs}) &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right)(s_{hy}^2 + \hat{\beta}_h^2 s_{hx}^2 - 2\hat{\beta}_h s_{hxy}) \\ &= \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) s_{hy}^2 (1 - r_h^2) \end{aligned}\right] \quad (8.15)$$

However, if a predetermined value β_{ho} is used for β_h, the bias B ($\bar{y}_{lrs}$) will be zero and $\hat{\beta}_h$ in mse ($\bar{y}_{lrs}$) will be replaced by β_{ho}. In this situation, the first expressions in (8.14) and (8.15) will not involve any approximation and will be exact. Also, these will not be equal to the corresponding second expressions, which will cease to hold in this case.

Example 8.4

A refrigerator manufacturing company contemplates to review its existing marketing policy. It has, therefore, decided to estimate the total number of refrigerators expected to be sold during the current summer season, half of which is almost over. Keeping various relevant factors in view, the whole country is divided into 4 zones. The number of registered dealers in these four zones are 400, 216, 364, and 274, whereas the total number of refrigerators sold by them during last summer are 29100, 12060, 26567, and 18111 respectively. Treating zones as strata, it was decided to select an overall sample of 42 dealers. Neyman allocation method was used to allocate it to different strata. The population mean squares for the number of refrigerators sold by a dealer during last summer season, for the four zones respectively, are 207.36, 282.24, 184.96, and 127.69. The samples of sizes 14, 9, 12, and 7 dealers were then selected from zones I, II, III,

and IV respectively (the procedure of allocation is explained in solution). The data in respect of number of refrigerators sold during last summer season and expected to be sold during current season are given in table 8.3 below.

Table 8.3 The number of refrigerators sold during last summer (LS) and expected sale for the current summer (CS)

Zone I		Zone II		Zone III		Zone IV	
LS (x)	CS (y)	LS (x)	CS (y)	LS (x)	CS (y)	LS (x)	CS (y)
53	69	44	52	60	67	58	52
84	80	67	73	76	86	65	71
93	87	84	78	78	75	56	62
66	72	52	60	68	77	48	44
77	81	48	42	55	64	73	77
82	94	62	56	48	45	85	80
68	64	56	50	86	98	61	66
84	88	70	76	91	95		
79	72	40	48	69	76		
98	110			70	79		
50	62			79	92		
78	70			49	66		
92	85						
63	77						

Solution

We are given that

$$n = 42,\ N_1 = 400,\ N_2 = 216,\ N_3 = 364,\ N_4 = 274,\ N = \sum_{h=1}^{4} N_h = 1254,$$

$$S_1^2 = 207.36,\ S_2^2 = 282.24,\ S_3^2 = 184.96,\ S_4^2 = 127.69,\ \text{and } X = \sum_{h=1}^{4} X_h = 85838.$$

The first step is to work out the sizes of samples to be selected from different strata. For this we need

$$\sum_{h=1}^{4} N_h S_h = (400)(14.4) + (216)(16.8) + (364)(13.6) + (274)(11.3)$$

$$= 17435.4$$

Then according to Neyman allocation method given in (5.14), the sample size for the h-th stratum is given by

$$n_h = n \frac{N_h S_h}{\Sigma N_h S_h}, \quad h = 1, 2, 3, 4$$

This gives

$$n_1 = (42) \frac{(400)(14.4)}{17435.4} = 13.9 \approx 14$$

$$n_2 = (42) \frac{(216)(16.8)}{17435.4} = 8.7 \approx 9$$

$$n_3 = (42) \frac{(364)(13.6)}{17435.4} = 11.9 \approx 12$$

$$n_4 = (42) \frac{(274)(11.3)}{17435.4} = 7.4 \approx 7$$

As in chapters 5 and 7, the sample means, sample mean squares, and products for each of the four strata have been computed, and are given in table 8.4 along with certain other sample values.

Table 8.4 Different population and sample values for the four strata

Zone I		Zone II		Zone III		Zone IV	
n_1	= 14	n_2	= 9	n_3	= 12	n_4	= 7
N_1	= 400	N_2	= 216	N_3	= 364	N_4	= 274
W_1	= .3190	W_2	= .1722	W_3	= .2903	W_4	= .2185
X_1	= 29100	X_2	= 12060	X_3	= 26567	X_4	= 18111
$\bar{X}_1$	= 72.8	$\bar{X}_2$	= 55.8	$\bar{X}_3$	= 73.0	$\bar{X}_4$	= 66.1
$\bar{y}_1$	= 79.4	$\bar{y}_2$	= 59.4	$\bar{y}_3$	= 76.7	$\bar{y}_4$	= 64.6
$\bar{x}_1$	= 76.2	$\bar{x}_2$	= 58.1	$\bar{x}_3$	= 69.1	$\bar{x}_4$	= 63.7
s^2_{1y}	= 166.7	s^2_{2y}	= 174.3	s^2_{3y}	= 226.6	s^2_{4y}	= 170.6
s^2_{1x}	= 211.1	s^2_{2x}	= 197.1	s^2_{3x}	= 193.0	s^2_{4x}	= 147.9
s_{1xy}	= 146.5	s_{2xy}	= 164.8	s_{3xy}	= 188.3	s_{4xy}	= 142.8
$\hat{\beta}_1$	= .6940	$\hat{\beta}_2$	= .8361	$\hat{\beta}_3$	= .9756	$\hat{\beta}_4$	= .9655
r_1	= .7810	r_2	= .8891	r_3	= .9004	r_4	= .8990

Since the β_h, h = 1, 2, 3, 4, values are likely to differ from stratum to stratum, we go for separate regression estimate for estimating the total expected sale of refrigerators. We, therefore, work out estimator $\hat{Y}_{lrs} = N\bar{y}_{lrs}$ of population total, where the estimator $\bar{y}_{lrs}$ is defined in (8.12). This yields

$$\hat{Y}_{lrs} = N\bar{y}_{lrs} = N \sum_{h=1}^{L} W_h [\bar{y}_h + \hat{\beta}_h (\bar{X}_h - \bar{x}_h)]$$

$$= 1254 [.3190 \{79.4 + (.6940)(72.8 - 76.2)\}$$

$$+ .1722 \{59.4 + (.8361)(55.8 - 58.1)\}$$

$$+ .2903 \{76.7 + (.9756)(73.0 - 69.1)\}$$

$$+ .2185 \{64.6 + (.9655)(66.1 - 63.7)\}]$$

$$= 1254 (24.58 + 9.90 + 23.37 + 14.62)$$

$$= 90877.38$$

$$\approx 90877$$

For estimating mean square error of $\hat{Y}_{lrs}$, we use second version of (8.15). Thus,

$$\text{mse}(\hat{Y}_{lrs}) = N^2 \text{mse}(\bar{y}_{lrs}) = N^2 \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) s_{hy}^2 (1 - r_h^2)$$

$$= (1254)^2 \left[(.3190)^2 \left(\frac{400 - 14}{(400)(14)}\right)(166.7)\{1 - (.7810)^2\}\right.$$

$$+ (.1722)^2 \left(\frac{216 - 9}{(216)(9)}\right)(174.3)\{1 - (.8891)^2\}$$

$$+ (.2903)^2 \left(\frac{364 - 12}{(364)(12)}\right)(226.6)\{1 - (.9004)^2\}$$

$$\left. + (.2185)^2 \left(\frac{274 - 7}{(274)(7)}\right)(170.6)\{1 - (.8990)^2\}\right]$$

$$= (1254)^2 (.4561 + .1153 + .2913 + .2175)$$

$$= 1698631.7$$

Below we work out the confidence limits in which the total number of refrigerators, expected to be sold during the current summer season, is likely to fall. These limits are given by

$$\hat{Y}_{lrs} \pm 2\sqrt{\text{mse}(\hat{Y}_{lrs})}$$

$$= 90877.38 \pm 2\sqrt{1698631.7}$$

$$= 90877.38 \pm 2606.63$$

$$= 88270.75, 93484.01$$

$$\approx 88271, 93484$$

In case, the information on refrigerators sold during the last summer season was not available, or not used, the expected total sale of refrigerators for this summer would have been estimated by using stratified simple random sampling WOR estimator given in (5.1), as

$$\hat{Y}_{st} = N \sum_{h=1}^{L} W_h \bar{y}_h = \sum_{h=1}^{L} N_h \bar{y}_h$$

The estimated variance of the estimator $\hat{Y}_{st}$, following (5.3), would have been

$$\begin{aligned} v(\hat{Y}_{st}) &= N^2 \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) s_{hy}^2 \\ &= \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h} s_{hy}^2 \\ &= \frac{400\,(400-14)}{14}(166.7) + \frac{216\,(216-9)}{9}(174.3) \\ &\quad + \frac{364\,(364-12)}{12}(226.6) + \frac{274\,(274-7)}{7}(170.6) \\ &= 6906833.9 \end{aligned}$$

Hence, the estimated percent relative efficiency of the separate linear regression estimator $\hat{Y}_{lrs}$ with respect to the stratified SRS without replacement estimator $\hat{Y}_{st}$, is given by

$$\begin{aligned} RE &= \frac{v(\hat{Y}_{st})}{mse\,(\hat{Y}_{lrs})}(100) \\ &= \frac{6906833.9}{1698631.7}(100) \\ &= 406.6 \end{aligned}$$

The increased efficiency of estimation can, therefore, be attributed to the use of auxiliary information through separate linear regression estimator. ■

When there is evidence that $\{\beta_h\}$ values do not differ much from stratum to stratum, we can use an alternative estimator $\bar{y}_{lrc}$, which is termed as *combined regression estimator*. Once again we assume that the sample sizes $\{n_h\}$ are sufficiently large, so that, the approximate expressions given here are close to the actual values. Now define

$$\beta_c = \frac{\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) S_{hxy}}{\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) S_{hx}^2}$$

$$\hat{\beta}_c = \frac{\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) s_{hxy}}{\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) s_{hx}^2} \tag{8.16}$$

Then we have the results (8.17) through (8.20).

Combined regression estimator of population mean $\bar{Y}$:

$$\left.\begin{aligned} \bar{y}_{lrc} &= \bar{y}_{st} + \hat{\beta}_c (\bar{X} - \bar{x}_{st}) \\ &= \sum_{h=1}^{L} W_h \bar{y}_h + \hat{\beta}_c (\bar{X} - \sum_{h=1}^{L} W_h \bar{x}_h) \end{aligned}\right] \tag{8.17}$$

Bias of estimator $\bar{y}_{lrc}$:

$$B(\bar{y}_{lrc}) = - \text{Cov}(\bar{x}_{st}, \hat{\beta}_c) \tag{8.18}$$

Approximate mean square error of estimator $\bar{y}_{lrc}$:

$$MSE(\bar{y}_{lrc}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) (S_{hy}^2 + \beta_c^2 S_{hx}^2 - 2\beta_c S_{hxy}) \tag{8.19}$$

Estimator of $MSE(\bar{y}_{lrc})$:

$$mse(\bar{y}_{lrc}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) (s_{hy}^2 + \hat{\beta}_c^2 s_{hx}^2 - 2\hat{\beta}_c s_{hxy}) \tag{8.20}$$

The bias in the separate regression estimator, as compared to the combined estimator, is large if the sample sizes $\{n_h\}$ are rather small. As stated earlier, if the regression coefficients $\{\beta_h\}$ do not seem to vary appreciably from stratum to stratum, a combined regression estimator should be preferred. However, if $\{\beta_h\}$ do vary from stratum to stratum, one should go for the separate estimator.

Example 8.5

In order to arrive at a more logical estimate of average leaf area for the newly developed strain of wheat (example 8.2), the physiologist raised 40 plants at each of the 3 different locations. This gave rise to 640, 710, and 769 leaves respectively. The total weight of these leaves, recorded at three locations, was found to be 69.000, 81.137, and 78.009 gm respectively. A WOR stratified random sample of 39 leaves was selected using proportional allocation (details given in solution). The observations recorded on sample leaves, in respect of area (in sq cm) and weight (in mg), are given in table 8.5. Estimate the average leaf area, and also build up the confidence interval for it.

Table 8.5 Area and weight for sample leaves

	Locations					
	I		II		III	
Leaf	Area (y)	Weight (x)	Area (y)	Weight (x)	Area (y)	Weight (x)
1	21.08	97	41.07	130	26.01	103
2	25.70	103	26.13	107	18.00	89
3	34.23	119	28.05	109	17.92	91
4	26.16	107	33.71	117	26.73	105
5	19.37	99	28.56	112	24.81	101
6	28.00	103	29.43	110	28.30	107
7	24.03	91	22.41	105	16.07	81
8	36.61	123	32.06	113	29.41	111
9	34.09	117	27.64	108	21.09	104
10	19.84	96	21.00	102	35.47	121
11	22.18	102	34.78	122	31.57	113
12	17.76	84	23.17	106	39.06	129
13			28.21	101	20.66	99
14					26.70	106

Solution

The number of units at three locations (strata) are $N_1 = 640$, $N_2=710$, and $N_3=769$, so that, $N=N_1+N_2+N_3 = 640+710+769 = 2119$. The total sample size is n=39 leaves. The number of units to be selected from each stratum are calculated by using (5.9). This gives

$$n_1 = \left(\frac{n}{N}\right)N_1 = \left(\frac{39}{2119}\right)640 = 11.8 \approx 12$$

$$n_2 = \left(\frac{n}{N}\right)N_2 = \left(\frac{39}{2119}\right)710 = 13.1 \approx 13$$

$$n_3 = \left(\frac{n}{N}\right)N_3 = \left(\frac{39}{2119}\right)769 = 14.2 \approx 14$$

The observations on area and weight from these selected leaves are given in table 8.5. The sample estimates $\bar{y}_h$, $\bar{x}_h$, s^2_{hy}, s^2_{hx}, and s_{hxy}, h=1,2,3, can be computed as in chapter 7. The estimates of regression coefficients are obtained from (8.11). The necessary computations are presented in table 8.6.

Table 8.6 Computations for samples selected from different strata

Location I	Location II	Location III
$n_1 = 12$	$n_2 = 13$	$n_3 = 14$
$N_1 = 640$	$N_2 = 710$	$N_3 = 769$
$W_1 = .3020$	$W_2 = .3351$	$W_3 = .3629$
$X_1 = 69000$	$X_2 = 81137$	$X_3 = 78009$
$\bar{y}_1 = 25.75$	$\bar{y}_2 = 28.94$	$\bar{y}_3 = 25.84$
$\bar{x}_1 = 103.4$	$\bar{x}_2 = 110.9$	$\bar{x}_3 = 104.3$
$\bar{X}_1 = 107.8$	$\bar{X}_2 = 114.3$	$\bar{X}_3 = 101.4$
$s^2_{1y} = 40.148$	$s^2_{2y} = 30.334$	$s^2_{3y} = 45.446$
$s^2_{1x} = 133.90$	$s^2_{2x} = 66.24$	$s^2_{3x} = 154.99$
$s_{1xy} = 68.410$	$s_{2xy} = 41.758$	$s_{3xy} = 81.065$
$\hat{\beta}_1 = .5109$	$\hat{\beta}_2 = .6304$	$\hat{\beta}_3 = .5230$
$r_1 = .933$	$r_2 = .932$	$r_3 = .966$

Calculations in table 8.6 indicate that the regression coefficients $\{\beta_h\}$ do not differ much from stratum to stratum. In such a situation, the combined regression estimator is more appropriate. We first work out stratified estimates of means for variables y and x. Thus from (5.1)

$$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h\bar{y}_h = \sum_{h=1}^{L} W_h\bar{y}_h$$

$$= (.3020)(25.75) + (.3351)(28.94) + (.3629)(25.84)$$

$$= 26.85$$

Similarly,

$$\bar{x}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h\bar{x}_h = \sum_{h=1}^{L} W_h\bar{x}_h$$

$$= (.3020)(103.4) + (.3351)(110.9) + (.3629)(104.3)$$

$$= 106.24$$

Also,

$$\sum_{h=1}^{L} W_h^2\left(\frac{N_h - n_h}{N_h\, n_h}\right)s_{hxy} = (.3020)^2\left(\frac{640-12}{(640)(12)}\right)(68.410)$$

$$+ (.3351)^2\left(\frac{710-13}{(710)(13)}\right)(41.758)$$

$$+ (.3629)^2\left(\frac{769-14}{(769)(14)}\right)(81.065)$$

$$= .5102 + .3541 + .7487$$

$$= 1.6130$$

$$\sum_{h=1}^{L} W_h^2\left(\frac{N_h - n_h}{N_h\, n_h}\right)s^2_{hx} = (.3020)^2\left(\frac{640-12}{(640)(12)}\right)(133.90)$$

$$+ (.3351)^2 \left(\frac{710-13}{(710)\ (13)}\right)(66.24)$$

$$+ (.3629)^2 \left(\frac{769-14}{(769)\ (14)}\right)(154.99)$$

$$= .9986 + .5617 + 1.4314$$

$$= 2.9917$$

Then on using the above computed values in (8.16), we obtain the estimate of combined regression coefficient β_c as

$$\hat{\beta}_c = \frac{1.6130}{2.9917} = .5392$$

Now,

$$X = 69000 + 81137 + 78009 = 228146$$

$$\overline{X} = \frac{228146}{2119} = 107.67$$

Using combined regression estimator defined in (8.17), the estimate of average leaf area is

$$\overline{y}_{lrc} = \overline{y}_{st} + \hat{\beta}_c (\overline{X} - \overline{x}_{st})$$

$$= 26.85 + (.5392)(107.67 - 106.24)$$

$$= 27.62$$

From (8.20), the estimate of mean square error of $\overline{y}_{lrc}$ is obtained as

$$\text{mse}(\overline{y}_{lrc}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h\ n_h}\right)(s_{hy}^2 + \hat{\beta}_c^2 s_{hx}^2 - 2\hat{\beta}_c s_{hxy})$$

$$= (.3020)^2 \left(\frac{640-12}{(640)\ (12)}\right)[40.148 + (.5392)^2(133.90) - 2(.5392)(68.410)]$$

$$+ (.3351)^2 \left(\frac{710-13}{(710)\ (13)}\right)[30.334 + (.5392)^2(66.24) - 2(.5392)(41.758)]$$

$$+ (.3629)^2 \left(\frac{769-14}{(769)\ (14)}\right)[45.446 + (.5392)^2(154.99) - 2(.5392)(81.065)]$$

$$= .0396 + .0387 + .0285$$

$$= .1068$$

The confidence interval for the actual average leaf area is now obtained from

$$\overline{y}_{lrc} \pm 2\sqrt{\text{mse}(\overline{y}_{lrc})}$$

$$= 27.62 \pm 2\sqrt{.1068}$$

$$= 27.62 \pm .65$$

$$= 26.97,\ 28.27$$

If the information on the leaf weight was not available, or it was not used, the average leaf area could be estimated by using the stratified SRS without replacement estimator $\bar{y}_{st}$ defined in (5.1). The estimator of variance for $\bar{y}_{st}$ is given by (5.3), which in this case becomes

$$v(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) s_{hy}^2$$

$$= (.3020)^2 \left(\frac{640-12}{(640)(12)}\right)(40.148) + (.3351)^2 \left(\frac{710-13}{(710)(13)}\right)(30.334)$$

$$+ (.3629)^2 \left(\frac{769-14}{(769)(14)}\right)(45.446)$$

$$= .2994 + .2572 + .4197$$

$$= .9763$$

Therefore, the estimated percent relative efficiency of the combined regression estimator, in relation to the stratified without replacement SRS estimator, is given by

$$RE = \frac{v(\bar{y}_{st})}{mse\,(\bar{y}_{lrc})}(100)$$

$$= \frac{.9763}{.1068}(100)$$

$$= 914.14$$

Hence, the use of information on leaf weight through the combined regression estimator has reduced the error of estimation for average leaf area to about one ninth of the stratified estimator which does not use any auxiliary information. ■

8.6 SOME FURTHER REMARKS

8.1 *Unbiased regression estimators* have been developed by Mickey (1959) and Williams (1963). Singh and Srivastava (1980) have proposed a sampling scheme for which the usual regression estimator becomes unbiased. Rao (1969) has found Mickey's estimator usually less efficient as compared to the usual regression estimator, in natural populations.

8.2 Often, the data are available on several auxiliary characteristics. In such cases, it will be beneficial to build up regression estimator which uses all the available information. Ghosh (1947) has proposed an estimator of the type

$$\bar{y}_{lrg} = \bar{y} + \sum_{i=1}^{k} \hat{\beta}_i (\bar{X}_i - \bar{x}_i)$$

where the symbols stand for their usual meaning, and k is the number of auxiliary variables on which the information is available.

8.3 Des Raj (1965) has proposed a *multivariate difference estimator* as

$$\bar{y}_{lrd} = \sum_{i=1}^{k} W_i [\bar{y} + \tau_i (\bar{X}_i - \bar{x}_i)]$$

where the weights W_i add up to unity and τ_i's are known constants.

LET US DO

8.1 Describe difference and regression methods of estimation for estimating population mean/total.

8.2 Using the data for a hypothetical population given in example 7.2, work out the relative efficiency of regression estimator $\bar{y}_{lr}$ with respect to the usual mean estimator $\bar{y}$, for a WOR simple random sample of size 3.

8.3 The total number of households in a development block consisting of 150 villages is 4066. A social scientist, interested in estimating total number of TV sets in the block, selected 30 villages using SRS without replacement procedure. Information on number of households and the number of TV sets for the 30 sample villages is given below :

Village	House-holds	TV sets	Village	House-holds	TV sets	Village	House-holds	TV sets
1	500	158	11	170	58	21	560	201
2	206	60	12	110	37	22	410	137
3	373	107	13	280	76	23	380	121
4	120	35	14	440	138	24	109	33
5	470	135	15	95	42	25	406	160
6	310	108	16	396	128	26	220	66
7	425	138	17	333	147	27	380	117
8	610	198	18	178	52	28	310	122
9	204	60	19	270	87	29	580	213
10	370	116	20	343	108	30	76	22

Estimate total number of TV sets in the block by using difference estimator, and also build up confidence interval for it. From an earlier survey, the value of the regression coefficient is known to be .35.

8.4 A campaign was launched to make the people aware of the fact that overweight persons run a heavy risk to their lives. As a result, 1000 overweight women got registered with a *Yoga Ashram* (an institution that conducts yoga lessons) for the purpose of reducing their weight and become more fit. At the time of registration, weight (in kg) of each woman was recorded and the average weight

of all the 1000 women came to be 65 kg. All the registered women were given lessons on the physical exercises to be undertaken daily. After 3 months, to determine the impact of the weight reducing program, the organizers selected a sample of 30 women and weighed them individually again. Present weight (y) and the initial weight (x) for the sample women are presented in the following table:

Woman	x	y	Woman	x	y	Woman	x	y
1	67.5	66.0	11	68.7	66.9	21	73.2	69.8
2	70.6	67.3	12	66.4	64.1	22	67.6	66.3
3	60.4	58.7	13	70.1	68.7	23	68.0	65.8
4	68.9	66.3	14	65.5	63.2	24	64.3	60.6
5	72.0	69.1	15	72.8	69.6	25	69.0	67.4
6	66.3	64.8	16	63.5	61.0	26	70.8	67.3
7	69.7	67.4	17	66.7	64.7	27	63.5	60.5
8	64.8	63.5	18	69.1	67.8	28	68.1	65.0
9	71.6	68.2	19	64.3	63.5	29	62.8	60.0
10	65.9	63.6	20	70.6	69.1	30	64.6	63.2

Using regression method of estimation, estimate the average present weight of a woman, and also build up the confidence interval for it. Also, compute its estimated relative efficiency in relation to the usual SRS estimator $\bar{y}$.

8.5 Employing regression method, estimate the total number of dwellings occupied by renters from the survey data given in exercise 7.6. Also, obtain confidence limits for it.

8.6 Using data of exercise 8.3, estimate the total number of TV sets in the block, and place confidence limits on it. Assume that the value of regression coefficient is not available in advance.

8.7 Suppose you are to estimate population total using regression method of estimation. Discuss how would you determine the optimal sample size from the information provided by a preliminary sample of size n_1?

8.8 Suppose that the sample of 30 villages, drawn from a total of 150 villages in exercise 8.3, is a preliminary sample. Using data of this exercise, examine whether this sample size is sufficient, or it has to be supplemented by selecting additional units, if one is interested in estimating total number of TV sets in the block with a margin of error equal to 50 TV sets ?

8.9 Assume that the sample of 30 women, drawn in exercise 8.4, is the preliminary sample. Using the data for this sample, comment whether, or not, this sample size is sufficient if the present average weight is to be estimated with a margin of error of 2 kg ? If not, what will you suggest ?

8.10 A plant breeder had limited amount of seed for 3 newly developed strains of sugarcane. He raised 36, 72, and 42 plants of strains 1, 2, and 3 respectively. Being

in possession of only a limited quantity of seed for the valuable strains, he could not afford to crush all the 150 canes to estimate its juice content. Using proportional allocation, he selected 6, 12, and 7 plants of strains 1, 2, and 3 respectively through simple random sampling WOR, so that, the overall sample was of 25 plants. Assume that the total weight of all the 150 canes is 70 kg. The quantity of juice and the weight of respective selected canes, both in grams, are given below :

Strain 1		Strain 2				Strain 3	
Cane	Juice	Cane	Juice	Cane	Juice	Cane	Juice
300	125	270	90	320	100	250	80
450	150	320	105	340	110	320	90
360	130	410	135	310	100	300	70
340	135	360	110	260	80	310	75
400	140	290	90	280	80	420	100
350	130	270	95	300	95	340	80
						280	70

Estimate juice quantity per cane by using separate regression estimator, and also determine the lower and upper confidence limits for it.

8.11 The Department of Animal Husbandry of a state government has undertaken a project on feeding and management practices of milch cows in a district comprising of 3 development blocks. These blocks, consisting of 70, 120, and 50 villages respectively, were treated as strata. A WOR random sample of 24 villages was drawn using proportional allocation. That means,

$$n_1 = \left(\frac{24}{240}\right) 70 = 7$$

$$n_2 = \left(\frac{24}{240}\right) 120 = 12$$

$$n_3 = \left(\frac{24}{240}\right) 50 = 5$$

villages were selected from strata I, II, and III respectively. The following table presents the total number of milch cows in 7, 12, and 5 randomly selected villages of strata I, II, and III respectively, during March 1993 and as per 1990 livestock census.

Stratum I		Stratum II				Stratum III	
1990	1993	1990	1993	1990	1993	1990	1993
17	21	21	18	18	24	16	21
19	22	14	21	8	15	21	18
9	11	16	20	16	11	13	16
13	14	18	24	26	21	19	25
22	18	26	20	11	16	20	23
16	21	13	20				
11	15	20	25				

From the 1990 census records, total number of milch cows in strata I, II, and III were 1260, 2400, and 1150 respectively. Using combined regression estimator, estimate the total number of milch cows in the district in March 1993. Also, obtain lower and upper confidence limits for it.

CHAPTER 9

Two-Phase Sampling

9.1 NEED FOR TWO-PHASE SAMPLING

The discussion in some of the previous chapters has revealed that the prior information on an auxiliary variable could be used to enhance the precision of an estimator. Ratio, product, and regression estimators require the knowledge of population mean $\bar{X}$ (or equivalently of total X) for the auxiliary variable x. For stratifying the population on the basis of the auxiliary variable, knowledge of its frequency distribution is required. When such information is lacking, it is some times less expensive to select a large sample (called *first-phase sample* or *initial sample*) on which auxiliary variable alone is observed. The purpose of this is to furnish a good estimate of $\bar{X}$, or equivalently of X. Frequency distribution for the auxiliary variable can also be estimated from observations made on the first-phase sample. A subsample (also called *final sample* or *second-phase sample*) from the initial sample is selected for observing the variable of interest. Information collected on the two samples is then used to construct estimators for the parameter under consideration.

As an illustration, let us consider the problem of estimating total production of cow milk in a certain region. For this purpose, we take village as the sampling unit and the number of milch cows in a village as the auxiliary variable. Since the total number of milch cows in all the villages of the region may not be available, the investigator could decide to take a large initial sample of villages, and collect information on number of milch cows in the sample villages. This information is then used to build up an estimate of X, the total number of milch cows in the region. The estimate of X, so obtained, could then be used in place of X in the ratio or regression estimate of total production of cow milk in the region. A subsample of villages is selected from the first-phase sample to observe the study variable, viz., cow milk yield in the village.

As yet another illustration, let us consider the situation of example 8.2, where a physiologist wanted to estimate average leaf area for a new strain of wheat. It may, sometimes, not be desirable to pluck all the leaves in the population of 120 plants and obtain total weight X for building a regression estimator of the average leaf area. It will, therefore, be more appropriate to select a large first-phase sample of leaves and measure weight for the sample leaves. A subsample from this initial sample of leaves could then be selected to determine leaf area. An estimate for the average weight $\bar{X}$ of leaves from

all the 120 plants could then be obtained from the observations made on the initial sample. This estimate can then be used in place of $\bar{X}$ in the regression estimate of the average leaf area.

The main point of deviation from the previously discussed procedures is that the sample is now drawn in two phases - first a large initial sample and then a subsample from this initial sample. Hence the name of the procedure.

Definition 9.1 *Two-phase sampling* (or *double sampling*) is a procedure where the lacking information on the auxiliary variable is collected from a large first-phase sample, and the study variable is observed on a smaller subsample selected from the first-phase sample.

When the sampling procedure is completed in three or more phases, the sampling procedure is termed as *multiphase sampling*. This procedure differs from the multistage sampling in the sense that the former requires a complete sampling frame of the ultimate sampling units, whereas in the latter a frame of the next stage units is necessary only for the sample units selected at that stage. Also, the sampling unit in case of multiphase sampling remains same at each phase of sampling, whereas it changes in case of multistage sampling. For example, if the sampling unit is a household, then in two-phase sampling, both first and second phase samples will be samples of households, whereas in case of multistage sampling if the first-stage sample is a sample of villages, the second-stage sample may be the sample of households.

It should be noted that in case of two-phase sampling, the size of the second sample on which we observe the study variable, will be reduced for the fixed total budget. This is because we had to spend part of the budget to observe auxiliary variable on the initial sample which could have otherwise been used to observe study variable on a comparatively larger sample. The technique is, therefore, beneficial only if the gain in precision is more than the loss in precision due to the reduction in the size of the final sample.

Two-phase sampling has been used in different ways by research workers. We shall, however, restrict our discussion on double sampling to ratio, product, regression, and PPS estimation procedures only.

9.2 TWO-PHASE SAMPLING IN RATIO, PRODUCT, AND REGRESSION METHODS OF ESTIMATION

While estimating mean/total through ratio, product, and regression methods of estimation, it was assumed that the population mean $\bar{X}$, or total X, for the auxiliary variable is known. If this information is lacking, the technique of double sampling provides an alternative. A large first-phase sample of n' units is drawn to estimate $\bar{X}$. Then, a final sample of n units is taken from n' units of the first-phase sample to observe the estimation variable y. If both the first-phase and the final samples are drawn using WOR equal probabilities

sampling, the different estimators and the other related results are given in sections 9.2.1, 9.2.2, and 9.2.3.

9.2.1 *Double Sampling for Ratio Method of Estimation*

Estimator of population mean $\bar{Y}$:

$$\bar{y}_{rd} = \frac{\bar{y}}{\bar{x}}\,\bar{x}' \tag{9.1}$$

where $\bar{x}' = (\Sigma x_i)/n'$, $i = 1, 2, \ldots n'$, is the mean for auxiliary variable x based on the initial sample of size n'.

Approximate bias of estimator $\bar{y}_{rd}$:

$$B(\bar{y}_{rd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{1}{\bar{X}}\right)(RS_x^2 - S_{xy}) \tag{9.2}$$

Approximate mean square error of estimator $\bar{y}_{rd}$:

$$MSE(\bar{y}_{rd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)(S_y^2 + R^2 S_x^2 - 2RS_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 \tag{9.3}$$

Estimator of MSE($\bar{y}_{rd}$) :

$$mse(\bar{y}_{rd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}\, s_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 \tag{9.4}$$

The symbols used have the same meaning as in chapters 7 and 8.

Example 9.1

The Sugar Mills Association of the state of Uttar Pradesh (India) wanted to estimate total man-hours lost due to strikes, power failure, breakdowns, etc., during February, 1992. The number of employees in a mill was taken as the auxiliary variable. The population total for this variable was, however, not available at the central office. A first-phase simple random WOR sample of 51 mills was, therefore, drawn from the population of 671 mills. Number of employees for each of these selected mills were recorded. A subsample of 24 mills was drawn from this initial sample using the same sampling scheme, and the man-hours lost were observed for the units included in the subsample. The information gathered on these two characters, for the units in the two samples, is given in the following table.

Table 9.1 Number of employees (x) and total man-hours lost (y) in '000 hour units

Mill	y	x	Mill	y	x	Mill	y	x
1		279	18	30.780	552	35	55.771	982
2	16.711	366	19		270	36		716
3		791	20	29.900	661	37		658
4	9.419	180	21		350	38	34.104	790
5		1001	22		690	39		381
6	14.370	291	23		570	40		280
7	20.609	420	24		240	41	21.680	411
8		371	25	41.460	691	42	9.413	150
9		687	26		524	43		398
10	9.358	196	27	41.220	832	44	11.639	241
11		792	28	10.004	266	45		336
12	52.024	1146	29		441	46		619
13		351	30		395	47	20.580	403
14		460	31	66.786	1246	48		186
15	20.113	370	32		179	49	36.009	864
16	13.200	220	33	18.632	413	50	12.600	316
17		485	34	10.852	286	51		298

Estimate the parameter in question, and place confidence limits on its population value.

Solution

From the statement of the problem, we have N=671, n'=51, and n= 24. Using table 9.1, we first work out the total number of employees in the sugar mills included in the initial sample. Thus,

$$\begin{aligned} x' &= x_1 + x_2 + \ldots + x_{51} \\ &= 279 + 366 + \ldots + 298 \\ &= 25041 \end{aligned}$$

It gives

$$\bar{x}' = \frac{25041}{51} = 491$$

Further, from the final subsample of n=24 mills, we compute the following sample estimates. These intermediate estimates will be used later. The calculations are analogous to those in chapters 7 and 8. Thus,

$$\bar{y} = \frac{1}{24}(16.711 + 9.419 + \ldots + 12.600)$$
$$= 25.3014$$

$$\bar{x} = \frac{1}{24}(366 + 180 + \ldots + 316)$$
$$= 512.2$$

$$\hat{R} = \frac{25.3014}{512.2} = .0494$$

Also,

$$s_y^2 = \frac{1}{24-1}[(16.711)^2 + (9.419)^2 + \ldots + (12.600)^2 - 24(25.3014)^2]$$
$$= 266.9345$$

$$s_x^2 = \frac{1}{24-1}[(366)^2 + (180)^2 + \ldots + (316)^2 - 24(512.2)^2]$$
$$= 100067.21$$

$$s_{xy} = \frac{1}{24-1}[(366)(16.711) + (180)(9.419) + \ldots + (316)(12.600)$$
$$- 24(512.2)(25.3014)]$$
$$= 5044.3834$$

Then, the estimate of correlation coefficient will be

$$r = \frac{s_{xy}}{s_x s_y} = \frac{5044.3834}{(\sqrt{100067.21})(\sqrt{266.9345})} = .976$$

As the man-hours lost and the number of employees in a mill are highly positively correlated, we decide to use ratio method of estimation for estimating total man-hours lost in February, 1992. Thus from (9.1), we have

$$\hat{Y}_{rd} = N\bar{y}_{rd} = \frac{N\bar{y}\,\bar{x}'}{\bar{x}}$$
$$= \frac{(671)(25.3014)(491)}{512.2}$$
$$= 16274.550$$

The total man-hours lost are, therefore, estimated as 16274550.

The estimate of mean square error $MSE(\hat{Y}_{rd})$ is worked out by using (9.4). It yields

$$\begin{aligned}
mse(\hat{Y}_{rd}) &= N^2\, mse(\bar{y}_{rd}) \\
&= N^2\left[\left(\frac{1}{n} - \frac{1}{n'}\right)(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right)s_y^2\right] \\
&= (671)^2\left[\left(\frac{1}{24} - \frac{1}{51}\right)\{266.9345 + (.0494)^2\,(100067.21) \right. \\
&\qquad \left. - 2(.0494)\,(5044.3834)\} + \left(\frac{1}{51} - \frac{1}{671}\right)(266.9345)\right] \\
&= 126624.68 + 2177452.7 \\
&= 2304077.4
\end{aligned}$$

Making use of the $mse(\hat{Y}_{rd})$ computed above, the required confidence limits for the population total are worked out as

$$\begin{aligned}
&\hat{Y}_{rd} \pm 2\sqrt{mse(\hat{Y}_{rd})} \\
&= 16274.550 \pm 2\sqrt{2304077.4} \\
&= 16274.550 \pm 3035.838 \\
&= 13238.712,\ 19310.388
\end{aligned}$$

Thus the Association could infer that the total man-hours lost during February, 1992, for the entire population of 671 sugar mills, are likely to fall in the closed interval [13238.712, 19310.388] thousand hours. ■

9.2.2 *Double Sampling for Product Method of Estimation*

Estimator of population mean $\bar{Y}$:

$$\bar{y}_{pd} = \frac{\bar{y}\,\bar{x}}{\bar{x}'} \tag{9.5}$$

Approximate bias of estimator $\bar{y}_{pd}$:

$$B(\bar{y}_{pd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)\frac{S_{xy}}{\bar{X}} \tag{9.6}$$

Approximate mean square error of estimator $\bar{y}_{pd}$:

$$MSE(\bar{y}_{pd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)(S_y^2 + R^2 S_x^2 + 2RS_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 \tag{9.7}$$

Estimator of MSE($\bar{y}_{pd}$):

$$\text{mse}(\bar{y}_{pd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)(s_y^2 + \hat{R}^2 s_x^2 + 2\hat{R} s_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 \quad (9.8)$$

The symbols used have the same meaning as before.

Example 9.2

A graduate student of statistics was asked to estimate average time per week for which the undergraduate students of a certain university view television (TV). The overall grade point average (OGPA) of the students was taken as the auxiliary variable. As the investigator found it difficult to record OGPA of all the 1964 undergraduate students, a first-phase sample of 150 students was selected. The OGPA of the students included in this initial sample were recorded from their personal files in the Registrar's office. The average OGPA for the first-phase sample was computed as 2.870 (on 4.00 basis). A subsample of 36 students was then selected from the first-phase sample. The students selected in the subsample were contacted personally to find the total time for which they view television in a week. The data in respect of both the characters, for the units selected in the subsample, are given in table 9.2.

Table 9.2 The OGPA (x) and number of hours per week (y) devoted to TV viewing

Student	y	x	Student	y	x	Student	y	x
1	8	2.51	13	14	2.86	25	5	3.25
2	3	3.41	14	11	2.24	26	3	3.49
3	1	3.25	15	3	3.43	27	8	2.63
4	5	3.04	16	6	3.25	28	4	3.61
5	12	2.73	17	7	2.73	29	13	2.17
6	6	3.10	18	5	2.91	30	14	3.10
7	9	2.58	19	4	3.07	31	6	3.01
8	2	3.46	20	10	2.61	32	8	2.58
9	0	3.69	21	8	2.48	33	9	2.41
10	8	2.83	22	12	3.39	34	4	2.96
11	9	2.91	23	6	2.95	35	5	2.85
12	6	3.06	24	1	3.77	36	10	3.74

Solution

In this problem, we are given that N = 1964, n' =150, n = 36, and $\bar{x}'$ = 2.870. From table 9.2, we have

$$\bar{y} = \frac{1}{36}(8+3+\ldots+10)$$

$$= 6.806$$

$$\bar{x} = \frac{1}{36}(2.51+3.41+\ldots+3.74)$$

$$= 3.002$$

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{6.806}{3.002} = 2.267$$

$$s_y^2 = \frac{1}{36-1}[8^2+3^2+\ldots+10^2-36(6.806)^2]$$

$$= 13.5897$$

$$s_x^2 = \frac{1}{36-1}[(2.51)^2+(3.41)^2+\ldots+(3.74)^2-36(3.002)^2]$$

$$= .1762$$

$$s_{xy} = \frac{1}{36-1}[(2.51)(8)+(3.41)(3)+\ldots+(3.74)(10)-36(3.002)(6.806)]$$

$$= -.9114$$

Then, the estimated correlation coefficient will be

$$r = \frac{-.9114}{(\sqrt{.1762})(\sqrt{13.5897})} = -.589$$

Since OGPA of a student and the time devoted by him to viewing TV, are negatively correlated, we shall use double sampling product estimator. The estimate of mean is then worked out by using (9.5). Thus,

$$\bar{y}_{pd} = \frac{\bar{y}\,\bar{x}}{\bar{x}'}$$

$$= \frac{(6.806)(3.002)}{2.870}$$

$$= 7.119$$

From (9.8), the estimate of mean square error of $\bar{y}_{pd}$ is obtained as

$$mse(\bar{y}_{pd}) = \left(\frac{1}{n}-\frac{1}{n'}\right)(s_y^2+\hat{R}^2 s_x^2+2\hat{R}\,s_{xy})+\left(\frac{1}{n'}-\frac{1}{N}\right)s_y^2$$

$$= \left(\frac{1}{36}-\frac{1}{150}\right)[13.5897+(2.267)^2(.1762)+2(2.267)(-.9114)]$$

$$+\left(\frac{1}{150}-\frac{1}{1964}\right)(13.5897)$$

$$= .2188+.0837$$

$$= .3025$$

We now work out the confidence interval for population mean. This is defined by the limits

$$\bar{y}_{pd} \pm 2\sqrt{mse(\bar{y}_{pd})}$$

$$= 7.119 \pm 2\sqrt{.3025}$$

$$= 7.119 \pm 1.100$$

$$= 6.019, 8.219$$

To conclude, an undergraduate student devotes, on the average, 7.119 hours per week to TV viewing. Also, the student conducting the survey is reasonably sure that had the information been collected for all the 1964 students, the per week average TV viewing time would have taken a value in the range 6.019 to 8.219 hours. ■

9.2.3 *Double Sampling for Regression Method of Estimation*

Estimator of population mean $\bar{Y}$:

$$\bar{y}_{lrd} = \bar{y} + \hat{\beta}(\bar{x}' - \bar{x}) \tag{9.9}$$

where $\hat{\beta} = s_{xy}/s_x^2$ as defined in section 8.3.

Approximate bias of estimator $\bar{y}_{lrd}$:

$$B(\bar{y}_{lrd}) = -\beta\left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{\mu_{21}}{S_{xy}} - \frac{\mu_{30}}{S_x^2}\right) \tag{9.10}$$

where $\mu_{21} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2(Y_i - \bar{Y})$ and $\mu_{30} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^3$.

Approximate mean square error of estimator $\bar{y}_{lrd}$:

$$\left.\begin{aligned} MSE(\bar{y}_{lrd}) &= \left(\frac{1}{n} - \frac{1}{n'}\right)(S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 \\ &= \left(\frac{1}{n} - \frac{1}{n'}\right)S_y^2(1-\rho^2) + \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 \end{aligned}\right] \tag{9.11}$$

Estimator of MSE $(\bar{y}_{lrd})$:

$$mse(\bar{y}_{lrd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)s_y^2(1-r^2) + \left(\frac{1}{n'} - \frac{1}{N}\right)s_y^2 \tag{9.12}$$

Symbols above have their usual meaning.

Example 9.3

Assume that in example 8.2, it was not possible to pluck all the 2106 leaves for weighing. Thus a first-phase WOR random sample of 120 leaves was selected and the sampled leaves were then plucked. The total weight of these 120 leaves was recorded as 13,400 mg. From this initial sample, a subsample of 33 leaves was drawn using same sampling scheme. Let us suppose that this subsample is same as the one considered in example 8.2. Therefore, using the sample data of table 8.2, estimate the average leaf area for the target population, and also build up confidence interval for it.

Solution

We have in this case, n' = 120 and $\bar{x}'$ = 13,400/120 = 111.667. Certain other sample estimates have also been calculated in example 8.2. These are reproduced below :

$\bar{y} = 27.263$, $\bar{x} = 110.545$, $s_y^2 = 61.003$, $s_x^2 = 187.631$, $s_{xy} = 100.312$,

$r = .9376$, and $\hat{\beta} = .5346$.

Since only the first-phase sample mean is available in place of the population mean, we go for regression estimator using double sampling. The estimator of mean in this case is given by (9.9). Thus, the estimate of average leaf area is

$$\begin{aligned}\bar{y}_{lrd} &= \bar{y} + \hat{\beta}(\bar{x}' - \bar{x}) \\ &= 27.263 + .5346(111.667 - 110.545) \\ &= 27.863\end{aligned}$$

Estimator of mean square error of $\bar{y}_{lrd}$ is obtained by using (9.12). Therefore,

$$\begin{aligned}mse(\bar{y}_{lrd}) &= \left(\frac{1}{n} - \frac{1}{n'}\right) s_y^2 (1 - r^2) + \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 \\ &= \left(\frac{1}{33} - \frac{1}{120}\right)(61.003)\{1 - (.9376)^2\} + \left(\frac{1}{120} - \frac{1}{2106}\right)(61.003) \\ &= .6414\end{aligned}$$

The confidence interval in which average leaf area for the population under consideration is likely to fall, with probability approximately .95, is determined by the limits

$$\begin{aligned}&\bar{y}_{lrd} \pm 2\sqrt{mse(\bar{y}_{lrd})} \\ &= 27.863 \pm 2\sqrt{.6414} \\ &= 27.863 \pm 1.6017 \\ &= 26.261, 29.465 \blacksquare\end{aligned}$$

9.3 SAMPLE SIZE DETERMINATION FOR RATIO, PRODUCT, AND REGRESSION ESTIMATORS

Let the cost function for two-phase sampling be

$$C = c_o + cn + c' n' \tag{9.13}$$

where c_o is the overhead cost, and c and c' are the per unit costs for observing the study and auxiliary variables respectively. For the sake of simplicity, we assume the population to be large so that 1/N is negligibly small. On minimizing the cost for the fixed variance, one gets the optimum values of n' and n. The values of n' and n thus obtained, are given in (9.14) and (9.15) for ratio, product, and regression estimators. The relations (9.16) and (9.17) give alternative formulas for the regression estimator.

First-phase and second-phase sample sizes for estimating mean using ratio, product, and regression estimators:

$$n' = \frac{S_y^2 - A}{V_o}\left[1 + \left\{\left(\frac{c}{c'}\right)\left(\frac{A}{S_y^2 - A}\right)\right\}^{\frac{1}{2}}\right] \tag{9.14}$$

$$n = n'\left[\left(\frac{c'}{c}\right)\left(\frac{A}{S_y^2 - A}\right)\right]^{\frac{1}{2}} \tag{9.15}$$

where

$A = S_y^2 + R^2 S_x^2 - 2RS_{xy}$ (for ratio estimator)

$A = S_y^2 + R^2 S_x^2 + 2RS_{xy}$ (for product estimator)

$A = S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy}$ (for regression estimator)

Alternative formulas for regression estimator :

$$n' = \frac{\rho^2 S_y^2}{V_o}\left[1 + \left\{\frac{c(1-\rho^2)}{c'\,\rho^2}\right\}^{\frac{1}{2}}\right] \tag{9.16}$$

$$n = n'\left[\frac{c'\,(1-\rho^2)}{c\rho^2}\right]^{\frac{1}{2}} \tag{9.17}$$

where ρ is the population correlation coefficient between y and x.

In surveys, depending on the precision required, the value of the variance V_o is fixed in advance. The costs c and c' are also known. Using the guess values of parameters S_y^2, S_x^2, S_{xy}, R, and ρ, one can arrive at optimum n' and n. In case the guess values of above said parameters are not available, then as in the previous chapters, a preliminary sample of n_1 units is selected and estimates of these parameters are obtained from these n_1 observations. These estimates are then used in place of the parameters involved in (9.14) to (9.17).

However, if the margin of error is to be fixed in terms of permissible error B, instead of the variance, then V_o in the above expressions will be replaced by $B^2/4$.

The results (9.14) to (9.17) can also be used for the estimation of population total by taking $V_o = (1/N^2)$ times the value of the variance fixed for the estimator of population total, or $V_o = B^2/4N^2$ when the margin of error is specified in terms of the permissible error B.

Example 9.4

Assume that the subsample drawn in example 9.2, is the preliminary sample of size $n_1 = 36$ students drawn from the population of 1964 students. Taking the costs of collecting information on OGPA and the time for which the students view TV as \$ .20 and \$ 1.25 per student respectively, determine the required subsample and the first-phase sample sizes if the variance is fixed at .20.

Solution

Here we have $n_1 = 36$, c' = \$ .20, c = \$ 1.25, and $V_o = .20$. Since we are using the observations on a preliminary sample for the determination of sample size, let us recall the sample values computed in example 9.2. Using the symbols of preceding chapters, we, therefore, write

$$s_{y1}^2 = 13.5897,\ s_{x1}^2 = .1762,\ s_{xy1} = -.9114,\ \text{and}\ \hat{R}_1 = 2.267.$$

Then,

$$\begin{aligned} A_1 &= s_{y1}^2 + \hat{R}_i^2 s_{x1}^2 + 2\hat{R}_1 s_{xy1} \\ &= 13.5897 + (2.267)^2 (.1762) + 2(2.267)(-.9114) \\ &= 10.3630 \end{aligned}$$

The optimum value of n' is given by (9.14). Using the analogous preliminary sample based values, it can be written as

$$n' = \frac{s_{y1}^2 - A_1}{V_o}\left[1 + \left\{\left(\frac{c}{c'}\right)\left(\frac{A_1}{s_{y1}^2 - A_1}\right)\right\}^{\frac{1}{2}}\right]$$

Substituting numerical values for different terms, we get

$$n' = \frac{13.5897 - 10.3630}{.20}\left[1 + \left\{\left(\frac{1.25}{.20}\right)\left(\frac{10.3630}{13.5897 - 10.3630}\right)\right\}^{\frac{1}{2}}\right]$$

$$= 88.4$$

$$\approx 88$$

Optimum value of n is then worked out through sample analog of (9.15), based on n_1 units. Thus, on redenoting the terms, it can be put as

$$n = n'\left[\left(\frac{c'}{c}\right)\left(\frac{A_1}{s_{y1}^2 - A_1}\right)\right]^{\frac{1}{2}}$$

On making substitutions, one gets

$$n = (88.4)\left[\left(\frac{.20}{1.25}\right)\left(\frac{10.3630}{13.5897 - 10.3630}\right)\right]^{\frac{1}{2}}$$

$$= 63.4$$

$$\approx 63$$

Thus, for getting an initial sample of 88 students, the preliminary sample of 36 students should be augmented by another SRS without replacement sample of 88-36=52 students selected from the population of 1964-36=1928 students left after the selection of the preliminary sample. These newly selected 52 students will be observed for their OGPA. To get a second-phase sample of 63 students, a subsample of 63-36= 27 students will be selected from the 52 newly selected students. These 27 students will also be observed for their TV viewing time. ■

9.4 TWO-PHASE PPS SAMPLING

In PPS sampling, the sample units are drawn with probability proportional to the size measure x. If the information on x is lacking for the population units, one can opt for two-phase sampling procedure. A first-phase sample of n' units is drawn from the given population of N units using SRS without replacement. The auxiliary variable is observed on these n' units. From this first-phase sample, a subsample of n units is selected by PPS with replacement method. The study variable is then measured on the subsample units. For i = 1, 2,..., n', let

$$x' = \sum_{i=1}^{n'} x_i$$

$$p_i' = \frac{x_i}{x'} \tag{9.18}$$

We then have the results (9.19) through (9.21).

Unbiased estimator of population mean $\bar{Y}$:

$$\bar{y}_{dp} = \frac{1}{n'n} \sum_{i=1}^{n} \frac{y_i}{p_i'} \tag{9.19}$$

Variance of estimator $\bar{y}_{dp}$:

$$V(\bar{y}_{dp}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \frac{n'-1}{nn'N(N-1)} \sigma_z^2 \tag{9.20}$$

Estimator of variance $V(\bar{y}_{dp})$:

$$v(\bar{y}_{dp}) = \frac{1}{N(n-1)(n'-1)} \left[\frac{N-1}{nn'} \sum_{i=1}^{n} \frac{y_i^2}{p_i'^2} + \frac{(n-1)(N-n')}{nn'} \left(\sum_{i=1}^{n} \frac{y_i^2}{p_i'} \right) - \{N(n'+n-1) - nn'\} \bar{y}_{dp}^2 \right] \tag{9.21}$$

where S_y^2 has been defined in (7.2), and $\sigma_z^2 = \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y \right)^2$ with $P_i = X_i/X$.

Example 9.5

An investigator is interested in estimating the average total money spent in a year on all the important festivals by a family in a certain locality consisting of 671 households. It is felt that the amount spent by a household on a festival depends on the income of the family, which in turn is related to the market price of the house where the family lives. Thus, the eye estimated price of the house was taken as the auxiliary variable. Since the determination of total eye estimated price of all the houses of the locality was time consuming, two-phase sampling was used to select a sample of households. A first-phase WOR simple random sample of 40 households was drawn. The eye estimated price of houses in the first-phase sample was determined and it totalled to 4980 thousand rupees. A subsample of 18 households was drawn, using PPS with replacement sampling, from 40 households of the first-phase sample. The data collected for these 18 households in respect of the two characters is given in table 9.3 as follows :

Table 9.3 Annual expenditure (y) on festivals by a household and the eye estimated price (x) of the house

Household	y	x	p'_i	Household	y	x	p'_i
1	2.50	70	.0141	21	9.00	295	.0592
2		25		22	4.80	170	.0341
3	3.70	110	.0221	23		84	
4		105		24	4.00	135	.0271
5		66		25		62	
6		73		26		30	
7	1.00	34	.0068	27	6.60	192	.0386
8	5.00	140	.0281	28	2.30	76	.0153
9		96		29		59	
10	6.50	260	.0522	30		173	
11		117		31	3.50	90	.0181
12		47		32	7.00	286	.0574
13	4.10	136	.0273	33		241	
14		126		34	2.70	54	.0108
15	2.65	80	.0161	35		134	
16		64		36		28	
17		182		37		85	
18		96		38	6.80	261	.0524
19		119		39	3.40	103	.0207
20	11.00	410	.0823	40		66	
Total						4980	

The units for the variables y and x in the table are in '00 rupees and '000 rupees respectively.

Estimate the annual expenditure on festivals per family, and build up confidence interval for it.

Solution

The statement of the problem gives N = 671, n' = 40, n = 18, and $x' = \Sigma x_i = 4980$.

The selection probabilities p'_i for the units included in the subsample are computed, by using (9.18), as

$$p'_1 = \frac{70}{4980} = .0141$$

$$p'_2 = \frac{110}{4980} = .0221$$

$$\vdots$$

$$p'_{18} = \frac{103}{4980} = .0207$$

These probabilities are given in table 9.3. Then,

$$\sum_{i=1}^{n} \frac{y_i}{p'_i} = \frac{2.50}{.0141} + \frac{3.70}{.0221} + \ldots + \frac{3.40}{.0207} = 2863.72$$

$$\sum_{i=1}^{n} \left(\frac{y_i}{p'_i}\right)^2 = \left(\frac{2.50}{.0141}\right)^2 + \left(\frac{3.70}{.0221}\right)^2 + \ldots + \left(\frac{3.40}{.0207}\right)^2 = 470900.37$$

$$\sum_{i=1}^{n} \frac{y_i^2}{p'_i} = \frac{(2.50)^2}{.0141} + \frac{(3.70)^2}{.0221} + \ldots + \frac{(3.40)^2}{.0207} = 13185.92$$

These sample values will be used for working out the estimate of mean and also the variance estimate $v(\bar{y}_{dp})$. From (9.19), the estimate of the average amount spent annualy on festivals by a family is given as

$$\bar{y}_{dp} = \frac{1}{n' n} \sum_{i=1}^{n} \frac{y_i}{p'_i}$$

$$= \frac{2863.72}{(40)(18)}$$

$$= 3.9774$$

Now for computing the estimate of variance $V(\bar{y}_{dp})$, we use (9.21). Thus,

$$v(\bar{y}_{dp}) = \frac{1}{N(n-1)(n'-1)} \left[\frac{N-1}{nn'} \sum_{i=1}^{n} \frac{y_i^2}{p'^2_i} + \frac{(n-1)(N-n')}{nn'} \left(\sum_{i=1}^{n} \frac{y_i^2}{p'_i} \right) - \{N(n'+n-1) - nn'\} \bar{y}^2_{dp} \right]$$

$$= \frac{1}{(671)(18-1)(40-1)} \left[\frac{671-1}{(18)(40)} (470900.37) + \frac{(18-1)(671-40)}{(18)(40)} (13185.92) - \{671(40+18-1) - (18)(40)\} (3.9774)^2 \right]$$

$$= \frac{1}{(671)(18-1)(40-1)} (438198.95 + 196451.89 - 593666.28)$$

$$= .09213$$

The required confidence limits will then be given by

$$\bar{y}_{dp} \pm 2\sqrt{v(\bar{y}_{dp})}$$

$$= 3.9774 \pm 2\sqrt{.09213}$$

$$= 3.3703, 4.5845$$

The investigator is thus reasonably sure that had all the 671 households been examined, the per family annual expenditure on festivals would have taken a value in the range from 337.03 to 458.45 rupees. ■

For determining optimum sample sizes n and n' in two-phase PPS with replacement sampling, one can proceed in the usual manner. However, the expression for the estimator of variance involved is somewhat complicated, and it makes determination of optimum n and n' quite difficult. Keeping the level of the book in mind, this topic is, therefore, not considered.

9.5 SAMPLING ON TWO OCCASIONS

In case of populations that are changing fast, census at long and infrequent intervals are not of much use. In such cases, it is desirable that the population is sampled at annual, or even at shorter, intervals regularly. Similarly, one may have to resort to repeated sampling of populations for which several kinds of data are to be collected and published at regular intervals. The method of sampling of units from the same population on successive occasions is called *multiple sampling* or *successive sampling*. This kind of sampling involves selection of samples on different occasions such that they have none, some, or all, units common with samples selected on previous occasions. In such surveys, it is possible to use information collected on previous occasions to improve the efficiency of estimators for subsequent occasions.

While resorting to repeated surveys, the objective of the investigator might be the estimation of one, or more, of the following parameters:

1. Average of population means over occasions.
2. Change in population mean value from one occasion to the next.
3. The population mean for the current occasion.

For estimating the average of population means over different occasions, the most appropriate strategy is to draw a fresh sample on each occasion. If it is desired to estimate the change in population mean from one occasion to the next, the same sample should be retained through all occasions. In case, the interest is to estimate the population average on the most recent occasion, the retaining of the part of the sample over occasions provides efficient estimates as compared to the other alternatives. As the theory for more than two occasions becomes complicated, we shall restrict our discussion to two occasions only. The problem of estimation in cases (1) and (2) above, is theoretically straightforward. In this section, we shall, therefore, only consider the problem of estimating population mean on the current occasion. For this, we shall assume that sample size on the two occasions is same.

Consider a population of N units and assume that the size of the population remains same over occasions and only the value of the study variable for different population units may get changed from occasion to occasion. Suppose a WOR simple random sample of n units is drawn on the first occasion. Out of this sample, m randomly selected units are retained on the second occasion. This sample is then augmented by selecting u additional units from the remaining population of (N-n) units using equal probability WOR sampling, so that, n = m+u. For h = 1, 2, let

$\bar{Y}_2$ = the population mean on the second occasion
$\bar{y}_h$ = the sample mean based on n units observed on the h-th occasion
$\bar{y}_{mh}$= the sample mean based on m matched units observed on the h-th occasion

$\bar{y}_{uh}$ = the sample mean based on u units drawn afresh on the h-th occasion
S_2^2 = the population mean square for second occasion
s_{m2}^2 = the sample mean square based on m matched units drawn on the second occasion
s_{u2}^2 = the sample mean square based on u units drawn afresh on the second occasion
s_2^2 = the pooled mean square based on matched and unmatched sample mean squares for the second occasion
ρ = the correlation coefficient between the variate of the second occasion and of the first occasion in the population
r = estimate of ρ based on matched part of the sample
$\hat{\beta}$ = estimate of regression coefficient β for the regression of variate of the second occasion on the variate of the first occasion

For the sake of simplicity, we take

$$p = \frac{m}{n} \text{ and } q = \frac{u}{n}$$

Let us further define

$$\frac{1}{W_u} = \frac{s_{u2}^2}{u} \tag{9.22}$$

$$\frac{1}{W_m} = \frac{(1-r^2)\, s_{m2}^2}{m} + \frac{r^2 s_{m2}^2}{n} \tag{9.23}$$

Then, we present the estimator of population mean on the second occasion along with its variance and estimator of varinace. Thus for large population, we have the following :

Estimator of population mean $\bar{Y}_2$:

$$\bar{y}_2 = \left(\frac{1}{W_m + W_u}\right)[W_u\, \bar{y}_{u2} + W_m \{\bar{y}_{m2} + \hat{\beta}\,(\bar{y}_1 - \bar{y}_{m1})\}] \tag{9.24}$$

Approximate variance of estimator $\bar{y}_2$:

$$V(\bar{y}_2) = \left(\frac{1-\rho^2 q}{1-\rho^2 q^2}\right)\frac{S_2^2}{n} \tag{9.25}$$

An estimator of variance V $(\bar{y}_2)$:

$$v(\bar{y}_2) = \left(\frac{1-r^2 q}{1-r^2 q^2}\right)\frac{s_2^2}{n} \tag{9.26}$$

where $s_2^2 = [(m-1)\, s_{m2}^2 + (u-1)\, s_{u2}^2]/(n-2)$.

The optimum value of q that minimizes the variance $V(\bar{y}_2)$, can be derived as

$$q_o = \frac{1}{1+\sqrt{1-\rho^2}} \tag{9.27}$$

This gives the fraction of the sample on the first occasion to be replaced on the second occasion, so that, one may achieve maximum precision. On substituting optimal q from (9.27) in (9.25), the minimum variance works out to be

$$V_0(\bar{y}_2) = [1+\sqrt{1-\rho^2}]\,\frac{S_2^2}{2n} \tag{9.28}$$

The estimate of $V_0(\bar{y}_2)$ is obtained by replacing ρ and S_2^2 in (9.28) by r and s_2^2 respectively. For further details, the reader may refer to Cochran (1977).

Example 9.6

A WOR simple random sample of 25 professors was drawn from 528 professors of a university, during the financial year 1991-92, to estimate average amount of money spent in buying National Saving Certificates (NSC). The government decided that from the financial year 1992-93, the interest on the amount of NSC purchased will be deducted from the total income for income-tax calculations, whereas, before 1992-93 this interest was counted towards personal income. Of the 25 professors selected in 1991-92, 12 were retained which constituted the matched part of the sample for the survey to be undertaken during 1992-93. A fresh sample of 13 professors was drawn using SRS without replacement from the remaining 528-25 = 503 professors. The information collected on both the occasions is presented in table 9.4.

Table 9.4 Amount (in '000 rupees) of NSC purchased

Professor	Amount of NSC		Professor	Amount of NSC	
	1991-92	1992-93		1991-92	1992-93
1	7.05	5.40	20	.50	
2	.40	.50	21	5.80	
3	1.00	2.02	22	2.70	
4	1.70	2.20	23	1.14	
5	1.32	.80	24	8.60	
6	0	.30	25	1.70	
7	1.50	1.00	26		1.35
8	.92	.50	27		1.80
9	1.20	.80	28		.90
10	.34	.20	29		4.00
11	3.10	5.20	30		5.20
12	1.40	2.00	31		.78
13	.98		32		2.50
14	.67		33		1.40
15	1.12		34		.75
16	.45		35		7.20
17	2.10		36		2.60
18	.90		37		3.00
19	1.10		38		2.40

Estimate the average amount spent by a professor for the purchase of NSC during 1992-93, and place confidence limits on it.

Solution

We are given that N=528, n=25, m=12, and u=13. This gives

$$p = \frac{12}{25} = .48 \text{ and } q = \frac{13}{25} = .52$$

Next, we compute the following averages :

$$\bar{y}_{m1} = \frac{1}{12}(7.05 + .40 + \ldots + 1.40)$$
$$= 1.661$$

$$\bar{y}_{m2} = \frac{1}{12}(5.40 + .50 + \ldots + 2.00)$$
$$= 1.743$$

$$\bar{y}_{u2} = \frac{1}{13}(1.35 + 1.80 + \ldots + 2.40)$$
$$= 2.606$$

$$\bar{y}_{1} = \frac{1}{25}(7.05 + .40 + \ldots + 1.70)$$
$$= 1.908$$

For working out r, we need sample mean squares and sample mean product based on matched part of the sample on first and second occasions. We, therefore, have

$$s^2_{m1} = \frac{1}{12-1}[(7.05)^2 + (.40)^2 + \ldots + (1.40)^2 - (12)(1.661)^2]$$
$$= \frac{1}{12-1}[71.7169 - 12(1.661)^2]$$
$$= 3.5100$$

$$s^2_{m2} = \frac{1}{12-1}[(5.40)^2 + (.50)^2 + \ldots + (2.00)^2 - (12)(1.743)^2]$$
$$= \frac{1}{12-1}[72.0304 - 12(1.743)^2]$$
$$= 3.2340$$

$$s_{m12} = \frac{1}{12-1}[(7.05)(5.40) + (.40)(.50) + \ldots + (1.40)(2.00)$$
$$- (12)(1.661)(1.743)]$$

$$= \frac{1}{12-1}\,[66.9940 - 12(1.661)\,(1.743)]$$

$$= 2.9320$$

This yields

$$\hat{\beta} = \frac{s_{m12}}{s_{m1}^2} = \frac{2.9320}{3.5100} = .8353$$

$$r = \frac{s_{m12}}{\sqrt{s_{m1}^2\, s_{m2}^2}} = \frac{2.9320}{\sqrt{(3.5100)\,(3.2340)}} = .8702$$

For obtaining the weights W_u and W_m, we need the values of s_{u2}^2 and s_{m2}^2 respectively. The value of s_{m2}^2 has already been computed. The value of s_{u2}^2 is obtained as

$$s_{u2}^2 = \frac{1}{13-1}\,[(1.35)^2 + (1.80)^2 + \ldots + (2.40)^2 - 13(2.606)^2]$$

$$= \frac{1}{13-1}\,[131.6534 - 13\,(2.606)^2]$$

$$= 3.6139$$

Thus, from (9.22) and (9.23), it follows that

$$\frac{1}{W_u} = \frac{s_{u2}^2}{u} = \frac{3.6139}{13} = .2780$$

$$\frac{1}{W_m} = \left[\frac{1-r^2}{m} + \frac{r^2}{n}\right] s_{m2}^2$$

$$= \left[\frac{1-(.8702)^2}{12} + \frac{(.8702)^2}{25}\right](3.2340)$$

$$= .1634$$

Hence, (9.24) yields

$$\bar{y}_2 = \left(\frac{1}{6.1200+3.5971}\right)[(3.5971)\,(2.606) + (6.1200)\,\{1.743 + (.8353)$$

$$(1.908-1.661)\}]$$

$$= \frac{21.3110}{9.7208}$$

$$= 2.192$$

From the sample information, it is thus estimated that, on the average, each professor spent Rs 2192 for purchasing NSC in 1992-93.

We now obtain the estimate of variance $V(\bar{y}_2)$ from (9.26). For this, we first calculate s_2^2.

$$s_2^2 = \frac{1}{25-2}[(12-1)(3.2340) + (13-1)(3.6139)]$$
$$= 3.4322$$

Thus,

$$v(\bar{y}_2) = \left[\frac{1-(.8702)^2(.52)}{1-(.8702)^2(.52)^2}\right]\frac{3.4322}{25}$$
$$= .1047$$

The confidence limits for the mean on second occasion are given by

$$\bar{y}_2 \pm 2\sqrt{v(\bar{y}_2)}$$
$$= 2.192 \pm 2\sqrt{.1047}$$
$$= 2.192 \pm .647$$
$$= 1.545, 2.839$$

Hence, one can be reasonably sure that, on the average, each professor spent rupees 1545 to 2839 on purchasing NSC during 1992-93. ■

It may be pointed out that the unmatched part of the sample on second occasion can also be selected from the entire population of N units, or from entire population minus the matched part of the sample. Both these strategies, however, yield less efficient estimators of population mean on second occasion, as compared to the one discussed in this section. For details, reader may refer to Ghangurde and Rao (1969), Singh (1972), and Cochran (1977).

9.6 SOME FURTHER REMARKS

9.1 Double sampling with regression estimator has been extended by Khan and Tripathi (1967) to the case where k auxiliary variables are observed on both the samples, and the population mean is estimated using multiple linear regression of the estimation variable on these k auxiliary variables.

9.2 Neyman (1938) was the first to discuss two-phase sampling for stratification. In stratified sampling, exact knowledge of strata sizes may sometimes be lacking. For instance, if the strata are based on old census data which do not indicate true situation for the current survey, double sampling provides an alternative in such situations. A first-phase sample is selected to collect information on the auxiliary variable. On the basis of this information, one can construct the strata and also estimate the sizes of different strata. Srinath (1971) and Rao (1973) have discussed the utility of double sampling for dealing with the problem of optimum allocation. For details, the reader may refer to Cochran (1977), Singh and Chaudhary (1989), and Des Raj (1968).

9.3 The second-phase sample could also be taken, independently of the first-phase sample, from the whole population. This practice is followed when, for instance, information on auxiliary variable is available with one agency, and information on both the study and auxiliary variables has been collected on a small independent sample by another agency. The theory for this case has been discussed by Des Raj (1968), Singh and Chaudhary (1989), and Cochran (1977).

LET US DO

9.1 In what kind of situations does the use of double sampling become necessary ?

9.2 What are the negative features of double sampling? Discuss.

9.3 "The two-phase sampling is beneficial if the estimate based on the information provided by initial and final sample is more precise per unit cost than the one based on a sample for estimation variable alone." Comment.

9.4 Give expressions for estimator $\bar{y}_{rd}$ of population mean $\bar{Y}$ and the estimator for variance $V(\bar{y}_{rd})$ in case of double sampling for ratio method of estimation. Assume that both the samples are drawn using SRS without replacement.

9.5 A total of 300 pieces of barren land in a district were marked for making them cultivable by applying gypsum to them. After 5 years of continuous application of gypsum, it was decided to estimate the total area from these pieces that has been brought under cultivation. A WOR random sample of 48 pieces of barren land was drawn, and eye estimates of the area of these pieces (x) were made. A subsample of 16 of these pieces was selected, and the area brought under cultivation (y) was measured for all the pieces in the subsample. The areas recorded in hectares are given in the following table.

Land piece	x	y	Land piece	x	y	Land piece	x	y
1	2.5		17	3.5		33	2.5	
2	1.0	0.6	18	2.5	2.2	34	3.0	
3	3.0		19	4.0		35	2.5	
4	4.0		20	5.0		36	3.0	2.4
5	4.0		21	5.5	4.9	37	1.0	
6	8.0	8.3	22	4.0		38	6.0	6.5
7	4.5		23	4.5		39	1.5	1.3
8	4.0	3.7	24	1.0		40	3.5	
9	5.0		25	5.2	4.7	41	4.0	
10	3.5		26	.5		42	2.0	
11	4.0		27	2.5		43	5.0	
12	6.5	5.8	28	9.0	8.4	44	6.5	
13	5.5		29	1.8		45	11.0	10.0
14	3.0		30	7.0	6.5	46	7.0	6.8
15	4.5		31	5.5		47	1.0	
16	4.5	4.1	32	2.5		48	2.5	1.7

Using double sampling ratio estimator, estimate the total area of barren land that has come under cultivation. Work out standard error of your estimate, and also build up the confidence interval for the population value.

9.6 In case of double sampling with simple random sampling WOR used for drawing both the samples, explain, how will you obtain expressions for the estimator $\hat{Y}_{pd}$ of population total Y and the estimator for variance $V(\hat{Y}_{pd})$ from the corresponding expressions for the mean estimator $\bar{y}_{pd}$?

9.7 A small survey study was conducted to estimate the total amount of money spent, during a year, on medical treatment by the faculty and staff of a certain Indian university. It is known that better paid employees and their dependents visit doctor/hospital for treatment less frequently. The reason is possibly their smaller family size and better living conditions. As the computation of total monthly salary paid to all the 4000 employees of the university is cumbersome, a preliminary WOR simple random sample of n'=100 employees was drawn. The total of monthly salaries (x) for these selected employees worked out to be Rs. 3,00,000 . A subsample of n=30 employees was then drawn using SRS without replacement. The amount of money spent on treatment during the preceding year (y) was obtained for each of the thirty selected employees. The information thus collected on x and y is given in the following table in hundred rupee units.

Employee	x	y	Employee	x	y	Employee	x	y
1	22.00	4.16	11	16.90	13.81	21	56.70	2.20
2	70.00	3.84	12	44.70	4.62	22	72.20	0
3	16.50	9.84	13	31.85	3.14	23	26.70	6.09
4	33.14	4.07	14	25.13	4.15	24	62.60	1.70
5	45.50	3.26	15	66.07	3.15	25	48.08	3.44
6	30.10	2.80	16	28.10	4.10	26	17.05	10.80
7	14.18	12.00	17	55.10	2.08	27	25.35	4.50
8	62.00	.60	18	15.80	9.70	28	80.10	0
9	48.60	1.76	19	18.20	11.10	29	70.96	2.50
10	90.30	2.00	20	82.30	.90	30	24.30	8.18

Using an appropriate estimator, determine the total amount of money spent on medical treatment during the preceding year by all the university employees. Also, place confidence limits on the population value.

9.8 For the data considered in exercise 9.5, estimate the total barren land area that has come under cultivation. Use double sampling based regression estimator for the purpose. Also, construct the confidence interval for the population value.

9.9 In the problem considered in exercise 8.4, assume that all the 1000 overweight women had registered for *yoga* excercises by mail. Since it was difficult to weigh all the women who got registered, a preliminary sample of 120 women was drawn using WOR simple random sampling, and initial weight of each of these women was recorded. The average initial weight of women in the preliminary sample came out to be 63.7 kg. All the 1000 women were provided video tapes and other literature containing lessons on the physical exercises to be undertaken. In order to determine the impact of the weight reducing program, the organizers selected a subsample of 30 women from the preliminary sample after 3 months of yoga participation. Each of these women was again weighed individually. Present weight (y) and the initial weight (x) for the women included in the subsample are as given in the table of exercise 8.4. Using double sampling based regression estimator, estimate the present average weight of a woman participant. Also, place the confidence limits on the present average weight of 1000 women registered for *yoga*.

9.10 Discuss, how will you determine the required initial and final sample sizes for ratio estimator in case of double sampling for fixed variance V_0 ? Assume that c_0 is the overhead cost, and c and c' are per unit costs for observing the study and auxiliary variables respectively. The cost function is as given in (9.13).

9.11 Assume that the subsample of n_1=24 units drawn in example 9.1, is a preliminary SRS without replacement sample from the population of 671 mills. Different estimates obtained from this subsample in example 9.1 are, therefore, to be treated as preliminary sample estimates. Thus, $s_y^2 = s_{y1}^2 = 266.9345$, $s_x^2 = s_{x1}^2 = 100067.21$,

$s_{xy} = s_{xyl} = 5044.3834$, and $\hat{R} = \hat{R}_l = .0494$. Let the costs of collecting information on number of employees per mill and the man-hours lost be Rs 10 and Rs 50 respectively. Determine the required initial and final sample sizes if the variance is to be fixed at 24,00,000 square man-hours.

9.12 Give expressions for the estimator of population mean and the estimator for its variance in case of two-phase PPS sampling. From these expressions, how will you write corresponding expressions for the estimator of population total and its variance estimator ?

9.13 A survey was conducted during 1987 to estimate the cattle population of 800 villages comprising a district. For this purpose, a WOR simple random sample of 54 villages was drawn. Again in 1992, taking number of cattle as the size variable, it was decided to select a PPS wih replacement subsample of 22 villages from the 54 villages selected in 1987. The number of cattle recorded in both the surveys, for the selected villages, is given below :

Village	1987	1992	Village	1987	1992	Village	1987	1992
1	980		19	1477	1610	37	354	
2	219		20	644	790	38	708	
3	640	670	21	340		39	250	285
4	710	775	22	887		40	790	842
5	572		23	317		41	1015	1160
6	693		24	166		42	317	
7	344		25	590		43	209	
8	599	620	26	270	310	44	144	195
9	412		27	621		45	990	1060
10	226	340	28	471		46	716	
11	170	210	29	156		47	681	763
12	488		30	1286	1340	48	550	
13	376		31	890		49	880	935
14	110		32	1364	1492	50	660	
15	1280	1385	33	570		51	505	
16	200	281	34	403	490	52	302	407
17	1170		35	780		53	290	376
18	918		36	291		54	170	

Estimate the total cattle population for the year 1992, and also place confidence limits on it.

9.14 A WOR simple random sample of 22 villages was drawn in 1984 from a development block of 160 villages in the Bathinda district of southern Punjab, to estimate camel population in the block. It is known that introduction of agricultural mechanization has adversely affected the camel population. In a subsequent survey conducted in 1992, out of 22 villages selected in the previous survey 10 were

retained to constitute the matched part of the sample for the survey of 1992. A fresh sample of 12 villages was drawn, using SRS without replacement, from the 138 villages not selected in the 1984 survey. The information on camel population, collected on both the occasions, is given in the table below:

Village	Camel population		Village	Camel population	
	1984	1992		1984	1992
1	40	13	18	39	
2	65	27	19	31	
3	31	10	20	46	
4	21	8	21	59	
5	67	29	22	63	
6	56	19	23		17
7	27	8	24		21
8	44	13	25		9
9	58	18	26		26
10	49	15	27		28
11	36		28		7
12	55		29		13
13	29		30		23
14	42		31		16
15	31		32		6
16	19		33		27
17	57		34		11

Estimate the camel population of the block in 1992, and also place confidence limits on it.

CHAPTER 10

Cluster Sampling

10.1 INTRODUCTION

Let us consider a situation where a study is to be carried out regarding the indebtedness of farmers of a particular region. For this purpose, a sample of farmers is to be selected. In case, the list of all the farmers (frame) in the region is not available, a simple random sample or a systematic sample of farmers, can not be selected. Even if the frame of all farmers in the region was available, a simple random sample of farmers will result in the sampled units (farmers) being scattered all over the region. This will require a good amount of travel to reach all the selected farmers for collecting information about the indebtedness and will, therefore, involve a formidable amount of travel expenditure.

On the other hand, the list of all the villages of the region is usually available and the characteristics of farmers in one village do not differ appreciably from the farmers of the other village. Therefore, if a simple random sample of villages, which could be considered as groups or *clusters* of farmers, is selected and all the farmers in the selected villages are enumerated, then a considerably reduced number of villages will account for the given number of farmers to be selected in the sample. In the new set-up, the total travel expenditure will be reduced to a great extent as the investigator will be required to travel between the few sample villages only. Besides, contacting of various farmers in any selected village involves comparatively very small cost. Also, since the characteristics of farmers in one village of the region may not be much different from the characteristics of farmers of the other village, not much loss of total information in the sample is expected.

All these considerations point to the need of selecting groups (clusters) of units together, and examine all the units contained in the selected clusters. The importance of the idea is that once a cluster has been reached, the cost of surveying units within the cluster is negligible.

Definition 10.1 The *cluster sampling* consists of forming suitable clusters of contiguous population units, and surveying all the units in a sample of clusters selected according to an appropriate sampling scheme.

For a given total number of units in the sample, the cluster sampling is usually less efficient than sampling of individual units as the latter is likely to provide a better cross section of the population units than the former. This is because of the tendency of units in a cluster to be similar. Also, the efficiency of cluster sampling is likely to

decrease with increase in cluster size. However, it is operationally convenient and economical than sampling of individual units. In many practical situations, the loss in efficiency from the view point of sampling variance is likely to be balanced by the reduction in cost. Hence, because of its operational convenience and possible reduction in cost, the survey tasks in many situations are facilitated by using nonoverlapping and collectively exhaustive clusters of units.

The clusters are usually formed by grouping neighboring units, or units which can be conveniently surveyed together. The construction of clusters, however, differs from the optimal construction of strata. Strata are to be as homogeneous as possible within themselves and differ as much as possible from one another with respect to the characteristic under study, and the units within a stratum need not be geographically contiguous. Clusters, on the other hand, should be as heterogeneous as possible within and as alike as possible between themselves. In such situations, cluster sampling is likely to be more efficient than the usual simple random sampling of same number of units from the population. Once appropriate clusters have been specified, a frame that lists all clusters in the population must be prepared. Various sampling procedures, viz., simple random sampling, varying probability sampling, stratified sampling, or systematic sampling, can be applied to cluster sampling by treating the clusters as sampling units. The expressions for the estimator, its variance, and estimator of variance can, therefore, be written in a straightforward manner. However, in this chapter, we shall only consider selection of clusters using simple random sampling and PPS with replacement sampling, and illustrate the steps involved in calculations of estimates of mean, total, and proportion.

10.2 NOTATIONS

In order to facilitate the understanding of the text, we first acquaint the reader with the notations to be used in the chapter. Let

N = number of clusters in the population

n = number of clusters in the sample

M_i = number of units in the i-th cluster of the population

$M_o = \sum_{i=1}^{N} M_i$ = total number of units in the population

$\bar{M} = M_o/N$ = average number of units per cluster in the population

Y_{ij} = value of the character under study for the j-th unit in the i-th cluster, $j = 1, 2, ..., M_i$; $i = 1, 2, ..., N$

$Y_{i.} = \sum_{j=1}^{M_i} Y_{ij}$ = i - th cluster total

$Y.. = \sum_{i=1}^{N} Y_{i.}$ = total of y - values for all the M_o units in the population

$$\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij} = \text{per unit i - th cluster mean}$$

$$y_{i.} = \sum_{j=1}^{M_i} y_{ij} = \text{i - th sample cluster total}$$

$$\bar{y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} = \text{per unit i - th sample cluster mean}$$

$$\bar{y}_c = \frac{1}{n} \sum_{i=1}^{n} y_{i.} = \text{mean per cluster in the sample}$$

$$\bar{\bar{Y}}_N = \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_i = \text{mean of cluster means in the population}$$

$$\bar{\bar{Y}} = \frac{1}{M_o} \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij} = \text{mean per unit of the population}$$

$$\bar{Y}_c = Y../N = \text{population mean per cluster}$$

10.3 ESTIMATION OF MEAN USING SIMPLE RANDOM SAMPLING

In practice, usually the clusters are of unequal sizes. For instance, households which are groups of persons, villages which are groups of households, fish catching centers which consist of groups of boats, could be considered as clusters for the purpose of sampling. In this chapter, we shall consider three estimators of population mean and total. The important results concerning these estimators are mentioned, assuming that a WOR simple random sample of n clusters has been drawn from N clusters, and all the units of the sample clusters are observed for the study variable. The results related to WR case can be obtained as particular cases from the results presented below for WOR sampling. We first consider the problem of estimating population mean per unit.

10.3.1 *Estimator 1*

Unbiased estimator of population mean when M_o is known :

$$\bar{y}_{c1} = \frac{N}{nM_o} \sum_{i=1}^{n} M_i \bar{y}_i = \frac{1}{\bar{M}n} \sum_{i=1}^{n} y_{i.} \qquad (10.1)$$

Variance of estimator $\bar{y}_{c1}$:

$$V(\bar{y}_{c1}) = \left(\frac{N-n}{Nn\bar{M}^2}\right) \frac{1}{N-1} \sum_{i=1}^{N} (Y_{i.} - \bar{Y}_c)^2 \qquad (10.2)$$

Estimator of variance $V(\bar{y}_{cl})$:

$$v(\bar{y}_{cl}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{1}{n-1}\sum_{i=1}^{n}(y_{i.} - \bar{M}\,\bar{y}_{cl})^2$$

$$= \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{1}{n-1}\left[\sum_{i=1}^{n} y_{i.}^2 - n(\bar{M}\,\bar{y}_{cl})^2\right] \qquad (10.3)$$

If the clusters are selected using WR sampling, then fpc = (N-n)/(N-1) in relation (10.2) and the sampling fraction f = n/N in (10.3) are taken as 1 and 0 respectively to get the corresponding results for the with replacement case.

Example 10.1

The recommended dose of nitrogen for wheat crop is 120 kg per hectare. A survey project was undertaken by the Department of Agriculture with a view to estimate the amount of nitrogen actually applied by the farmers. For this purpose, 12 villages from a population of 170 villages of a development block were selected using equal probabilities WOR sampling, and the information regarding the nitrogen use was collected from all the farmers in the selected villages. The data collected are presented in table 10.1. The total number of farmers in these 170 villages is available from the *patwari*'s record as 2890. Estimate the average amount of nitrogen used in practice by a farmer. Also, obtain standard error of the estimate, and place confidence limits on the population mean.

Table 10.1 Per hectare nitrogen (in kg) applied to wheat crop by farmers

Village	M_i	Nitrogen applied (in kg) by a farmer									$y_{i.}$
1	15	105	128	130	108	135	122	120	138	126	1843
		117	125	126	123	118	122				
2	18	135	128	105	130	120	125	114	128	121	2206
		109	128	122	129	112	133	117	119	131	
3	25	124	118	128	106	132	121	126	108	136	3085
		121	128	125	136	128	121	127	122	113	
		117	132	128	125	130	109	124			
4	21	108	116	111	129	119	137	129	121	118	2582
		126	131	128	134	125	112	121	116	114	
		129	127	131							
5	11	114	105	126	132	116	125	104	121	132	1292
		106	111								
6	13	128	116	132	136	121	122	129	123	127	1627
		118	134	126	115						
7	22	103	118	107	128	132	136	124	129	130	2686
		134	108	106	117	129	113	118	126	127	
		129	119	125	128						

Table 10.1 continued...

Village	M_i	Nitrogen applied (in kg) by a farmer									$y_{i.}$
8	12	109	121	114	128	133	135	114	128	107	1471
		125	126	131							
9	10	119	128	117	131	105	128	136	113	127	1234
		130									
10	20	130	127	116	128	114	120	127	123	134	2449
		122	126	121	117	125	129	122	113	111	
		126	118								
11	10	126	117	124	121	131	133	126	120	128	1242
		116									
12	16	124	121	127	119	120	123	128	117	121	1935
		93	115	120	124	121	130	132			

Solution

Here we have N = 170, M_o = 2890, and n = 12. It gives

$$\overline{M} = \frac{M_o}{N} = \frac{2890}{170} = 17$$

As the sample cluster totals $y_{i.}$ will be required for computing the estimates, these are worked out below and are presented in the last column of the table 10.1 above.

Cluster 1 : $y_{1.} = 105 + 128 + \ldots + 122 = 1843$
Cluster 2 : $y_{2.} = 135 + 128 + \ldots + 131 = 2206$
. .
. .
. .
Cluster 12 : $y_{12.} = 124 + 121 + \ldots + 132 = 1935$

Estimate of the average amount of nitrogen used per hectare, by a farmer, follows from (10.1) as

$$\bar{y}_{c1} = \frac{1}{\overline{M}n} \sum_{i=1}^{n} y_{i.}$$

$$= \frac{1}{(17)(12)} (1843 + 2206 + \ldots + 1935)$$

$$= \frac{23652}{(17)(12)}$$

$$= 115.941$$

We then work out the estimate of variance using (10.3). Thus,

$$v(\bar{y}_{c1}) = \left(\frac{N-n}{Nn\bar{M}^2}\right) \frac{1}{n-1} \sum_{i=1}^{n} (y_{i.} - \bar{M}\,\bar{y}_{c1})^2$$

where

$$\bar{M}\,\bar{y}_{c1} = (17)\,(115.941) = 1970.997$$

Hence,

$$v(\bar{y}_{c1}) = \left(\frac{170-12}{(170)\,(12)\,(17)^2}\right) \frac{1}{11} [(1843-1970.997)^2 + (2206-1970.997)^2 + \ldots + (1935-1970.997)^2]$$

$$= \left(\frac{170-12}{(170)\,(12)\,(17)^2}\right) \frac{1}{11} [(1843)^2 + (2206)^2 + \ldots + (1935)^2 - 12(1970.997)^2]$$

$$= \frac{(170-12)\,(4331060)}{(170)\,(12)\,(17)^2(11)}$$

$$= 105.519$$

Using above calculated estimate of variance, the standard error of mean will be

$$se(\bar{y}_{c1}) = \sqrt{105.519}$$

$$= 10.272$$

Following (2.8), the required confidence limits for population mean are obtained as

$$\bar{y}_{c1} \pm 2\sqrt{v(\bar{y}_{c1})}$$

$$= 115.941 \pm 20.544$$

$$= 95.397,\ 136.485$$

The above confidence limits reasonably ensure that per hectare average dose of nitrogen used by a farmer in the target population is likely to be within the range 95.397 to 136.485 kg. ■

10.3.2 *Estimator 2*

The estimator $\bar{y}_{c1}$ in (10.1) for population mean assumes the knowledge of M_o. When M_o is not known and the values of M_i are known only for the sample clusters, then $\bar{Y}$ can be estimated by using the estimator $\bar{y}_{c2}$.

Estimator of population mean which does not depend on M_o:

$$\bar{y}_{c2} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i \tag{10.4}$$

Bias of the estimator $\bar{y}_{c2}$:

$$B(\bar{y}_{c2}) = -\frac{1}{\bar{M}} \text{Cov}(\bar{y}_i, M_i) \tag{10.5}$$

Variance of estimator $\bar{y}_{c2}$:

$$V(\bar{y}_{c2}) = \left(\frac{N-n}{Nn}\right) \frac{1}{N-1} \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y}_N)^2 \tag{10.6}$$

Estimator of variance $V(\bar{y}_{c2})$:

$$\begin{aligned} v(\bar{y}_{c2}) &= \left(\frac{N-n}{Nn}\right) \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{c2})^2 \\ &= \left(\frac{N-n}{Nn}\right) \frac{1}{n-1} \left(\sum_{i=1}^{n} \bar{y}_i^2 - n\bar{y}_{c2}^2\right) \end{aligned} \tag{10.7}$$

The bias for the estimator $\bar{y}_{c2}$, given by (10.5), is expected to be small when the cluster means $\bar{Y}_i$ and sizes M_i are not highly correlated. In such a case, it is advisable to use this estimator since the variance of the estimator $\bar{y}_{c2}$ is likely to be less than the variance $V(\bar{y}_{c1})$ given in (10.2). The bias of this estimator disappears if the cluster sizes $M_1, M_2, \ldots, M_N$ are equal.

10.3.3 *Estimator 3*

An alternative estimator of population mean $\bar{Y}$ for the situation where M_i's are known only for sample clusters (irrespective of whether M_o is known or not), is the ratio type estimator. This estimator is also biased but the bias decreases with increase in n.

Estimator of population mean which does not depend on M_o:

$$\bar{y}_{c3} = \frac{\sum_{i=1}^{n} y_{i.}}{\sum_{i=1}^{n} M_i} \tag{10.8}$$

Approximate bias of estimator $\bar{y}_{c3}$:

$$B(\bar{y}_{c3}) = \left(\frac{N-n}{Nn}\right) \frac{1}{\bar{M}^2} (\bar{Y} S_m^2 - S_{my}) \tag{10.9}$$

where S_m^2 and S_{my} are defined in (7.2) with cluster size m replacing x.

Approximate variance of estimator $\bar{y}_{c3}$:

$$V(\bar{y}_{c3}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{1}{N-1}\sum_{i=1}^{N}(Y_{i.}-\bar{Y}M_i)^2 \qquad (10.10)$$

Estimator of variance $V(\bar{y}_{c3})$:

$$v(\bar{y}_{c3}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{1}{n-1}\sum_{i=1}^{n}(y_{i.}-M_i\bar{y}_{c3})^2 \qquad (10.11)$$

$\bar{M}$ in (10.11) above can be replaced by $\hat{\bar{M}} = \frac{1}{n}\sum_{i=1}^{n} M_i$, if it is not already known.

The bias in the estimator $\bar{y}_{c3}$ becomes zero if the cluster sizes $M_1, M_2, ..., M_N$ are equal. The estimator $\bar{y}_{c3}$ is also expected to be more efficient than the estimator $\bar{y}_{c1}$ when M_i and $y_{i.}$ are highly positively correlated.

Example 10.2

A state government wanted to estimate the extent of tax evasion, per passenger, by the private bus owners on a certain route. Being the busy route, it was decided to check the buses at random. The total number of buses that leave the terminal daily is 80. The buses were serially numbered depending on the time of their departure. Fifteen buses were then selected with SRS without replacement. The tickets with all the passengers of the selected buses were examined enroute, and the amount of tax evasion was recorded. The total of passenger tax evaded for each sampled bus was then computed, and is given in table 10.2 along with the total number of passengers in the bus. Estimate the average tax evaded per passenger by the private bus operators, and place a confidence interval on the population average.

Table 10.2 Tax evaded (in rupees) per sampled bus along with cluster mean

Bus	Passengers (M_i)	Tax evaded ($y_{i.}$)	Cluster (bus) mean ($\bar{y}_i$)
1	60	118.70	1.98
2	70	148.30	2.12
3	65	140.10	2.16
4	52	98.40	1.89
5	72	109.50	1.52
6	48	72.05	1.50
7	54	100.20	1.86
8	60	115.10	1.92
9	43	108.70	2.53

Table 10.2 continued ...

Bus	Passengers (M_i)	Tax evaded ($y_{i.}$)	Cluster (bus) mean ($\bar{y}_i$)
10	69	135.45	1.96
11	58	117.30	2.02
12	74	150.70	2.04
13	55	126.40	2.30
14	69	95.30	1.38
15	66	111.65	1.69
Total	915	1747.85	28.87

Solution

In this problem, we have N = 80, n = 15, and M_o is not known. Although the choice between estimators 2 and 3 depends on the value of the correlation coefficient as mentioned earlier, but for the sake of illustration we demonstrate the use of both the estimators.

Use of estimator 2. The estimate of the tax evaded per passenger is obtained by using (10.4) as

$$\bar{y}_{c2} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i$$

$$= \frac{1}{15} (1.98 + 2.12 + \ldots + 1.69)$$

$$= \frac{28.87}{15}$$

$$= 1.92$$

We then compute the estimate of variance using (10.7).

$$v(\bar{y}_{c2}) = \left(\frac{N-n}{Nn}\right) \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{c2})^2$$

$$= \left(\frac{80-15}{(80)(15)}\right) \frac{1}{14} [(1.98-1.92)^2 + (2.12-1.92)^2 + \ldots + (1.69-1.92)^2]$$

$$= \left(\frac{80-15}{(80)(15)}\right) \frac{1}{14} [(1.98)^2 + (2.12)^2 + \ldots + (1.69)^2 - (15)(1.92)^2]$$

$$= \frac{(80-15)(1.5979)}{(80)(15)(14)}$$

$$= .006182$$

The confidence interval for the population average can be derived from

$$\bar{y}_{c2} \pm 2\sqrt{v(\bar{y}_{c2})}$$

$$= 1.92 \pm 2\sqrt{.006182}$$

$$= 1.92 \pm .16$$

$$= 1.76, 2.08$$

The confidence limits computed above, indicate that the daily per passenger evasion of tax by the population of private bus owners is likely to fall in the closed interval [1.76, 2.08] rupees.

Use of estimator 3. The estimate of per passenger tax evaded by the private bus operators is given by

$$\bar{y}_{c3} = \frac{\sum_{i=1}^{n} y_{i.}}{\sum_{i=1}^{n} M_i}$$

Substituting the values from table 10.2, one gets

$$\bar{y}_{c3} = \frac{1747.85}{915} = 1.91$$

For computing the estimate of variance, we use expression (10.11). Thus,

$$v(\bar{y}_{c3}) = \left(\frac{N-n}{Nn\bar{M}^2}\right) \frac{1}{n-1} \sum_{i=1}^{n} (y_{i.} - M_i \bar{y}_{c3})^2$$

Since $\bar{M}$ is unknown, we use its estimate $\hat{\bar{M}}$, where

$$\hat{\bar{M}} = \frac{1}{n} \sum_{i=1}^{n} M_i = \frac{915}{15} = 61$$

Then,

$$v(\bar{y}_{c3}) = \left(\frac{80-15}{(80)(15)(61)^2}\right) \frac{1}{14} [\{118.70 - (60)(1.91)\}^2$$

$$+ \{148.30 - (70)(1.91)\}^2 + \ldots + \{111.65 - (66)(1.91)\}^2]$$

$$= \frac{(80-15)(4509.041)}{(80)(15)(61)^2(14)}$$

$$= .004688$$

We now compute confidence limits for daily per passenger tax evasion by all the private bus operators on the route under consideration. These we find as

$$\bar{y}_{c3} \pm 2\sqrt{v(\bar{y}_{c3})}$$
$$= 1.91 \pm 2\sqrt{.004688}$$
$$= 1.91 \pm .14$$
$$= 1.77, 2.05$$

These confidence limits are very close to those obtained earlier by using estimator 2. ■

10.4 ESTIMATION OF TOTAL USING SIMPLE RANDOM SAMPLING

An estimator of population total can be easily obtained by multiplying any one of the corresponding estimators of mean given in (10.1), (10.4), and (10.8) by M_o.

Estimators of population total Y:

$$\hat{Y}_{c1} = \frac{N}{n}\sum_{i=1}^{n} y_{i.} \qquad (10.12)$$

$$\hat{Y}_{c2} = \frac{M_o}{n}\sum_{i=1}^{n} \bar{y}_i \qquad (10.13)$$

$$\hat{Y}_{c3} = \frac{M_o \sum_{i=1}^{n} y_{i.}}{\sum_{i=1}^{n} M_i} \qquad (10.14)$$

Expressions for variances and their estimators for the above estimators of population total, can be easily obtained by multiplying their counterparts for mean by M_o^2.

The estimator $\hat{Y}_{c1}$ in (10.12) can be used even when M_o is not known while the estimators $\hat{Y}_{c2}$ or $\hat{Y}_{c3}$ can be used only when M_o, or equivalently $\bar{M}$, is known. If the correlation between the cluster means and cluster size is low, the estimator $\hat{Y}_{c2}$ may be preferred. However, in case of large samples when correlation between the cluster total and cluster size is positive and high, use of estimator $\hat{Y}_{c3}$ is advised.

Example 10.3

Along the sea coast of an Indian state, there are 120 small villages. Some of the residents of these villages resort to fishing for their livelihood. The list of these villages is available. However, no information is available about the number of families (M_i) involved in the said profession in these villages. For estimating the total catch of fish by the villagers,

16 villages were selected using SRS without replacement. The information collected from all the families of sample villages about the catch of fish on a particular day, is given in table 10.3. Estimate total catch of fish on this day for the entire population of 120 villages. Also, build up confidence interval for the population total.

Table 10.3 Catch of fish (in quintals) per family for the selected villages

Village	M_i	Catch of fish by families										$y_{i.}$
1	11	6.8	5.0	2.4	7.5	3.2	4.4	4.0	3.0	6.8	5.6 9.2	57.9
2	7	4.3	5.9	6.0	1.7	2.3	7.3	2.6				30.1
3	8	2.0	4.1	6.6	7.0	8.2	9.9	1.4	1.2			40.4
4	6	5.3	6.8	2.9	3.3	4.8	1.8					24.9
5	4	1.8	10.5	1.0	2.7							16.0
6	10	8.4	1.5	6.7	8.8	5.6	2.8	1.2	4.5	6.8	8.1	54.4
7	7	5.5	2.8	4.6	1.0	10.9	6.7	3.8				35.3
8	8	5.4	6.4	3.3	8.1	7.4	6.5	1.9	2.0			41.0
9	7	6.0	7.1	2.1	1.8	6.7	8.3	2.9				34.9
10	6	2.8	3.4	6.9	1.1	6.6	4.8					25.6
11	3	5.6	6.8	1.6								14.0
12	5	9.9	6.4	1.2	2.0	3.6						23.1
13	9	8.5	2.0	1.8	4.9	6.7	5.2	7.7	5.9	1.6		44.3
14	8	1.2	4.5	6.8	1.0	2.3	3.7	6.9	8.1			34.5
15	7	3.8	6.4	7.6	2.5	6.1	3.5	1.9				31.8
16	8	2.6	10.0	6.3	9.8	1.0	1.8	7.8	2.7			42.0
Total												550.2

Solution

First, we work out sample cluster (which is village in this case) totals as

Cluster 1 : $y_{1.} = 6.8 + 5.0 + \ldots + 9.2 = 57.9$

Cluster 2 : $y_{2.} = 4.3 + 5.9 + \ldots + 2.6 = 30.1$

. .

. .

. .

Cluster 16 : $y_{16.} = 2.6 + 10.0 + \ldots + 2.7 = 42.0$

The cluster totals obtained this way, are presented in table 10.3.

Since M_o is not known, the estimate of total catch of fish is obtained by using (10.12). Thus,

$$\hat{Y}_{cl} = \frac{N}{n} \sum_{i=1}^{n} y_{i.}$$

$$= \frac{(120)\,(550.2)}{16}$$

$$= 4126.50$$

The estimate of variance of $\hat{Y}_{cl}$ can be worked out from the estimate of variance of mean given in (10.3), after multiplying it by M_o^2. This means

$$v(\hat{Y}_{cl}) = \frac{N(N-n)}{n\,(n-1)} \sum_{i=1}^{n} (y_{i.} - \overline{M}\,\bar{y}_{cl})^2$$

Also,

$$\overline{M}\,\bar{y}_{cl} = \frac{\hat{Y}_{cl}}{N} = \frac{4126.50}{120} = 34.3875$$

Hence,

$$v(\hat{Y}_{cl}) = \frac{(120)\,(120-16)}{(16)\,(15)} [(57.9-34.3875)^2 + (30.1-34.3875)^2$$

$$+ \ldots + (42.0-34.3875)^2]$$

$$= \frac{(120)\,(120-16)}{(16)\,(15)} [(57.9)^2 + (30.1)^2 + \ldots + (42.0)^2 - 16(34.3875)^2]$$

$$= \frac{(120)\,(120-16)\,(2263.9974)}{(16)\,(15)}$$

$$= 117727.86$$

The required confidence interval for population total is given by

$$\hat{Y}_{cl} \pm 2\sqrt{v(\hat{Y}_{cl})}$$

$$= 4126.50 \pm 2\sqrt{117727.86}$$

$$= 4126.50 \pm 686.23$$

$$= 3440.27,\ 4812.73$$

This means that the total catch of fish for this day, for all the 120 villages under study, is likely to take a value in the interval 3440.27 to 4812.73 quintals, with confidence coefficient as .95. ■

Example 10.4

A state is planning to set up a small spinning mill in a hilly area. Before finalizing the capacity, layout, etc., the administration thinks it appropriate to have information regarding the number of unemployed males/females that are educated up to at least 5th grade but have not attained the age of 45 years, in the villages falling within a radius of 15 km. The number of such villages is 60, and the total number of households, comprising these villages, is known to be 560. Twelve villages were selected using SRS without replacement method. The information collected from all the households in the sample villages, is given below in table 10.4.

Table 10.4 Unemployed persons below 45 years and educated up to at least 5th grade

Village	M_i	Number of unemployed males/females														$y_{i.}$	$\bar{y}_i$
1	9	1	2	0	1	2	1	1	3	2						13	1.4444
2	6	3	4	0	1	1	2									11	1.8333
3	7	2	4	1	2	2	1	1								13	1.8571
4	14	1	0	0	2	4	0	2	3	1	2	0	3	5	2	25	1.7857
5	8	2	1	0	1	1	0	4	2							11	1.3750
6	9	2	4	3	1	0	2	3	2	0						17	1.8889
7	8	1	1	5	0	2	1	2	1							13	1.6250
8	10	3	1	0	3	0	4	1	0	0	4					16	1.6000
9	11	2	0	3	4	0	1	0	2	3	1	0				16	1.4545
10	11	0	2	1	5	3	2	0	2	3	1	1				20	1.8182
11	9	0	1	3	5	0	3	1	1	1						15	1.6667
12	12	4	1	2	1	2	0	2	4	3	2	1	1			23	1.9167
Total	114															193	20.2655

Estimate the total number of unemployed persons in question, and construct confidence interval for this population total.

Solution

In this problem, the total number of households in the population of 60 villages is known, that is, $M_o = 560$. Also, $N = 60$ and $n = 12$. Since M_o is known, we can use both the estimators defined in (10.13) and (10.14). Although the magnitudes of the correlation coefficients between the pairs of variables $(\bar{y}_i, M_i)$ and $(y_{i.}, M_i)$ are to decide as to which of these two estimators is to be preferred in practice, we shall consider both for the purpose of illustration.

Use of estimator 2. As in the previous examples, we first work out cluster totals and cluster means. Thus,

$$y_{1.} = 1 + 2 + \ldots + 2 = 13, \qquad \bar{y}_1 = 13/9 = 1.4444$$
$$y_{2.} = 3 + 4 + \ldots + 2 = 11, \qquad \bar{y}_2 = 11/6 = 1.8333$$
$$\vdots \qquad\qquad \vdots$$
$$y_{12.} = 4 + 1 + \ldots + 1 = 23, \qquad \bar{y}_{12} = 23/12 = 1.9167$$

The cluster totals and means, so calculated, are presented in table 10.4.

The estimate of population total from (10.13) is given by

$$\hat{Y}_{c2} = \frac{M_o}{n} \sum_{i=1}^{n} \bar{y}_i$$

Using $\Sigma \bar{y}_i$, i = 1, 2, ..., n, computed in table (10.4), one gets the estimate of the total number of unemployed persons with required qualifications as

$$\hat{Y}_{c2} = \frac{560}{12} (20.2655) = 945.72 \approx 946$$

So far as the estimate of variance $V(\hat{Y}_{c2})$ is concerned, it can be computed from (10.7), after multiplying it by M_o^2. Thus we get

$$v(\hat{Y}_{c2}) = \frac{M_o^2 (N-n)}{Nn(n-1)} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{c2})^2$$

$$= \frac{M_o^2 (N-n)}{Nn(n-1)} \left(\sum_{i=1}^{n} \bar{y}_i^2 - n\bar{y}_{c2}^2 \right)$$

where

$$\bar{y}_{c2} = \frac{\hat{Y}_{c2}}{M_o} = \frac{945.72}{560} = 1.6888$$

This yields

$$v(\hat{Y}_{c2}) = \frac{(560)^2 (60-12)}{(60)(12)(11)} [(1.4444 - 1.6888)^2 + (1.8333 - 1.6888)^2$$
$$+ \ldots + (1.9167 - 1.6888)^2]$$
$$= \frac{(560)^2 (60-12)}{(60)(12)(11)} [(1.4444)^2 + (1.8333)^2 + \ldots + (1.9167)^2$$
$$- 12(1.6888)^2] = 746.03$$

Confidence interval for the total unemployed work force is then given by

$$\hat{Y}_{c2} \pm 2\sqrt{v(\hat{Y}_{c2})}$$

$$= 945.72 \pm 2\sqrt{746.03}$$

$$= 945.72 \pm 54.63$$

$$= 891.09, 1000.35$$

$$\approx 891, 1000$$

One can, therefore, say with probability approximately .95, that the total number of unemployed persons of the desired kind ranges from 891 to 1000.

Use of estimator 3. The estimate of the total number of unemployed persons with desired qualifications is now given by

$$\hat{Y}_{c3} = \frac{M_o \sum_{i=1}^{n} y_{i.}}{\sum_{i=1}^{n} M_i} = \frac{(560)(193)}{114} = 948.07 \approx 948$$

Estimate of variance $V(\hat{Y}_{c3})$ is worked out by multiplying (10.11) by M_o^2. For this, we calculate

$$\bar{y}_{c3} = \frac{\hat{Y}_{c3}}{M_o} = \frac{948.07}{560} = 1.6930$$

Then,

$$v(\hat{Y}_{c3}) = \frac{N(N-n)}{n(n-1)} \sum_{i=1}^{n} (y_{i.} - M_i \bar{y}_{c3})^2$$

$$= \frac{(60)(60-12)}{(12)(11)} [\{13 - 9(1.6930)\}^2 + \{11 - 6(1.6930)\}^2$$

$$+ \ldots + \{23 - 12(1.6930)\}^2]$$

$$= \frac{(60)(48)(35.4954)}{(12)(11)}$$

$$= 774.4451$$

As usual, the confidence interval for the total number of unemployed persons with specified qualifications is given by

$$\hat{Y}_{c3} \pm 2\sqrt{v(\hat{Y}_{c3})}$$

$$= 948.07 \pm 2\sqrt{774.4451}$$

$$= 948.07 \pm 55.66$$

$$= 892.41, 1003.73$$

$$\approx 892, 1004$$

Thus, the confidence intervals yielded by the two estimators are quite similar. ■

10.5. RELATIVE EFFICIENCY OF CLUSTER SAMPLING

In this section, we consider the relative efficiency aspect of cluster sampling. For this purpose, we assume the simplest situation where all the clusters in the population are of equal size, that is, $M_i = M$ for i = 1, 2, ..., N. In this case, the estimators of population mean given in (10.1) and (10.4) become identical, and hence, expressions for their variances and estimators of variances are also same. Therefore, whatever we conclude about estimator (10.1) will also hold for estimator in (10.4). In case of cluster sampling considered in this section, we select a sample of nM units in the form of n clusters each consisting of M units. Thus, if the same number of units were selected from the population of NM units by SRS without replacement, the simple mean estimator $\bar{y}$ and its variance will be given by

$$\bar{y} = \frac{1}{nM} \sum_{i=1}^{nM} y_i \tag{10.15}$$

$$V(\bar{y}) = \left(\frac{1}{nM} - \frac{1}{NM}\right) S^2$$

$$= \left(\frac{N-n}{NnM}\right) \frac{1}{NM-1} \left(\sum_{i=1}^{NM} Y_i^2 - NM\bar{Y}^2\right) \tag{10.16}$$

The relative efficiency of the estimator $\bar{y}_{cl}$ in (10.1), in relation to the simple mean estimator $\bar{y}$, will, therefore, be given by

$$RE = \frac{V(\bar{y})}{V(\bar{y}_{cl})} \tag{10.17}$$

where the variance $V(\bar{y}_{cl})$ is available in (10.2).

The relative efficiency, defined in (10.17), involves values of study variable for all population units. In practice, however, the investigator has only the sample observations on n clusters of M units each. Thus, he can only estimate the relative efficiency from the sample observations. For this, we shall need the estimates of two variances involved in (10.17). An unbiased estimator of $V(\bar{y})$, from a cluster sample, is given by

$$v_c(\bar{y}) = \frac{N-n}{(NM-1)n} \left[\frac{1}{nM} \sum_{i=1}^{n} \sum_{j=1}^{M} y_{ij}^2 + v(\bar{y}_{cl}) - \bar{y}_{cl}^2\right] \tag{10.18}$$

The estimate of relative efficiency in (10.17) will then be given as in (10.19).

The estimated RE of estimator $\bar{y}_{cl}$ with respect to the usual estimator $\bar{y}$, from a cluster sample :

$$RE = \frac{v_c(\bar{y})}{v(\bar{y}_{cl})} \tag{10.19}$$

where $v(\bar{y}_{cl})$ and $v_c(\bar{y})$ are given in (10.3) and (10.18) respectively.

Example 10.5

In a developing country, a certain company has 25 centers located at different places in a state. Each center has been provided with 4 telephones. A student attending a sample survey course was given an assignment to estimate the average number of calls per telephone made on a typical day for this company. The student did not have the telephone facility, and was also short of funds. Because of this, he selected 5 centers using SRS without replacement. The number of calls made on a typical working day from each telephone of the sample centers were recorded personally. The data so obtained are summarized in table 10.5.

Table 10.5 Number of calls made from selected centers

Center	M_i	Calls made				$y_{i.}$	$\bar{y}_i$
1	4	26	34	27	25	112	28
2	4	44	33	28	31	136	34
3	4	18	33	25	28	104	26
4	4	37	21	22	40	120	30
5	4	23	34	42	29	128	32

Estimate the average number of daily calls per telephone made from all the 25 centers, by using estimator (10.1). Also, estimate the relative efficiency of the estimator used with respect to the usual simple mean estimator, from the sample selected above.

Solution

Here $N = 25$, $n = 5$, and $M_i = M = 4$. The sample cluster means are given in the last column of table 10.5. The estimate of average number of daily calls is computed using estimator $\bar{y}_{c1}$ given in (10.1). Thus for $M_i = M$,

$$\bar{y}_{c1} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i \qquad \text{[same as } \bar{y}_{c2} \text{ in (10.4)]}$$

From the last column of table 10.5, we get

$$\bar{y}_{c1} = \frac{1}{5}(28+34+26+30+32)$$

$$= 30$$

For $M_i = M$, $i = 1, 2, \ldots, N$, the variance estimator $v(\bar{y}_{c1})$ in (10.3) becomes

$$v(\bar{y}_{c1}) = \frac{N-n}{Nn(n-1)} \left(\sum_{i=1}^{n} \bar{y}_i^2 - n\bar{y}_{c1}^2 \right)$$

On making substitutions, one gets

$$v(\bar{y}_{c1}) = \frac{25-5}{(25)(5)(4)} [(28)^2 + (34)^2 + \ldots + (32)^2 - 5(30)^2]$$

$$= 1.6$$

Now from (10.18), the variance estimator of the simple mean estimator $\bar{y}$ in (10.15), from the selected cluster sample, will be

$$v_c(\bar{y}) = \frac{N-n}{(NM-1)n}\left[\frac{1}{nM}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij}^2 + v(\bar{y}_{c1}) - \bar{y}_{c1}^2\right]$$

We first compute the term involving sum of squares of all the individual observations. Thus,

$$\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij}^2 = (26)^2 + (34)^2 + \ldots + (29)^2$$

$$= 18962$$

Then,

$$v_c(\bar{y}) = \frac{25-5}{[(25)(4)-1](5)}\left[\frac{18962}{(5)(4)} + 1.6 - (30)^2\right]$$

$$= 2.0081$$

On using (10.19), the estimate of percent relative efficiency will be

$$RE = \frac{2.0081}{1.6}(100)$$

$$= 125.5 \blacksquare$$

We know that the mean square error/variance of any estimator is related to the number of units selected in the sample. In the following section, we, therefore, consider the problem of determining the required number of clusters to be included in the sample when one is to estimate the population mean or total with specified amount of tolerable error in the estimate.

10.6 DETERMINING THE SAMPLE SIZE FOR ESTIMATING MEAN/TOTAL

The total volume of information in cluster sampling is affected by two factors, namely, the number of clusters in the sample and the cluster size. Assuming that the cluster size has been fixed in advance, we consider the problem of determining the number of clusters required to be selected in the sample to obtain estimators with a given precision. Let a preliminary sample of n_1 clusters be selected initially. Based on the information obtained from these n_1 clusters, we compute for estimator 1

$$s_{c1}^2 = \frac{1}{(n_1-1)\overline{M}^2}\sum_{i=1}^{n_1}(y_{i.} - \overline{M}\,\bar{y}_{c1})^2$$

where

$$\overline{M}\,\bar{y}_{c1} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{i.}$$

Using s_{c1}^2 in place of

$$\frac{1}{(n-1)\overline{M}^2}\sum_{i=1}^{n}(y_{i.}-\overline{M}\,\overline{y}_{c1})^2$$

in (10.3), we solve the equation

$$2\sqrt{v(\overline{y}_{c1})} = B$$

or equivalently

$$2\sqrt{\frac{N-n}{Nn}s_{c1}^2} = B$$

for n. This gives us the required number of clusters to be selected in the overall sample. Here B is the half width of the confidence interval for population mean, and represents the error which the investigator is willing to tolerate in the estimate for population mean. Similarly, we can also obtain the formulas for the required sample size in case of other two estimators. All these formulas are listed below :

Sample size required to estimate mean/total with B as tolerable error :

$$n = \frac{Ns_{ci}^2}{ND+s_{ci}^2},\ i=1,2,3 \qquad (10.20)$$

where D and s_{ci}^2 for the three estimators are defined as follows:

Estimator 1 :

$$s_{c1}^2 = \frac{1}{\overline{M}^2(n_1-1)}\sum_{i=1}^{n_1}(y_{i.}-\overline{M}\,\overline{y}_{c1})^2 \qquad (10.21)$$

$$D = \frac{B^2}{4} \quad \text{(when estimating mean)}$$

$$D = \frac{B^2}{4M_o^2} \quad \text{(when estimating total)}$$

Estimator 2 :

$$s_{c2}^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(\overline{y}_i-\overline{y}_{c2})^2 \qquad (10.22)$$

$$D = \frac{B^2}{4} \quad \text{(when estimating mean)}$$

$$D = \frac{B^2}{4M_o^2} \quad \text{(when estimating total)}$$

Estimator 3 :

$$s_{c3}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_{i.} - M_i \bar{y}_{c3})^2 \qquad (10.23)$$

$$D = \frac{B^2 \bar{M}^2}{4} \qquad \text{(when estimating mean)}$$

$$D = \frac{B^2}{4N^2} \qquad \text{(when estimating total)}$$

In all above cases, the estimators $\bar{y}_{ci}$, i = 1, 2, 3, are based on the preliminary sample, and $\bar{M}$, if unknown, is also to be estimated from the preliminary sample. If $n_1 \geq n$, then preliminary sample is sufficient, otherwise, $(n-n_1)$ additional clusters are to be selected to get the required overall sample.

Example 10.6

Suppose that the information of example 10.1 pertains to a preliminary sample of 12 villages. Using the data of this example, verify whether this sample size is sufficient to make the inference about the average amount of nitrogen used per hectare by farmers, with a permissible error of magnitude 12 kg ?

Solution

We have N = 170, M_o = 2890, and the preliminary sample size n_1=12. Further, from (10.21)

$$s_{c1}^2 = \frac{1}{\bar{M}^2 (n_1 - 1)} \sum_{i=1}^{n_1} (y_{i.} - \bar{M} \bar{y}_{c1})^2$$

Using information from intermediate computations for the estimate of variance in example 10.1, one gets

$$s_{c1}^2 = \frac{4331060}{(17)^2 (11)}$$

$$= 1362.397$$

Also,

$$D = \frac{B^2}{4} = \frac{(12)^2}{4} = 36$$

Then the sample size required to estimate the average amount of nitrogen used per hectare with a bound of 12 kg on the error of estimation, would be obtained by using (10.20). Thus,

$$n = \frac{N s_{cl}^2}{ND + s_{cl}^2}$$

$$= \frac{(170)(1362.397)}{(170)(36) + 1362.397}$$

$$= 30.95$$

$$\approx 31$$

This shows that to draw the inference about the average dose of nitrogen actually used in practice by the farmers, with a tolerable error of 12 kg, the preliminary sample size of 12 villages is not sufficient. The investigator will need to select 31-12=19 more villages to achieve the required degree of precision. ■

10.7 ESTIMATION OF PROPORTION

Sometimes the investigator is interested in estimating proportion of units in a population, belonging to a specified category (say) A. For instance, he/she may wish to estimate proportion of heads of religious places who are graduates, or the proportion of persons who drink. On defining Y_{ij} as 1 if the j-th unit of i-th cluster belongs to the specified class, and 0 otherwise, $Y_{i.}$ yields the total number of units in the i-th cluster that belong to category A. Let a_i denote the number of such units in cluster i (we shall denote the corresponding random variable by a), then the population mean per unit reduces to the population proportion P, since

$$\bar{Y} = \frac{1}{M_o} \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}$$

$$= \frac{M_A}{M_o}$$

$$= P$$

where $M_A = \Sigma a_i$, i = 1, 2, ..., N, denotes the total number of units in the population that belong to category A.

The estimators of proportion of units belonging to category A corresponding to the estimators $\bar{y}_{ci}$, i = 1, 2, 3, expressions of their variances, and estimators of variances can be obtained in a straightforward manner by replacing $y_{i.}$ by a_i and $\bar{Y}$ by P in (10.1) to (10.11).

10.7.1 *Estimator 1*

The reader should note that this estimator of population proportion can only be used when M_o, the total number of units in the population, is known.

Unbiased estimator of population proportion when M_o is known :

$$p_{c1} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} a_i \tag{10.24}$$

Variance of estimator p_{c1}:

$$V(p_{c1}) = \frac{N-n}{Nn(N-1)} \sum_{i=1}^{N} \left(\frac{a_i}{\bar{M}} - P\right)^2 \tag{10.25}$$

Estimator of variance $V(p_{c1})$:

$$v(p_{c1}) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} \left(\frac{a_i}{\bar{M}} - p_{c1}\right)^2 \tag{10.26}$$

10.7.2 *Estimator 2*

This estimator can be used in both the cases where M_o is known, or when it is not known.

Estimator of population proportion which does not depend on M_o :

$$p_{c2} = \frac{1}{n} \sum_{i=1}^{n} \frac{a_i}{M_i} \tag{10.27}$$

Bias of the estimator p_{c2} :

$$B(p_{c2}) = -\frac{1}{\bar{M}} \text{Cov}\left(\frac{a_i}{M_i}, M_i\right) \tag{10.28}$$

Variance of estimator p_{c2} :

$$V(p_{c2}) = \frac{N-n}{Nn(N-1)} \sum_{i=1}^{N} \left(\frac{a_i}{M_i} - \bar{P}\right)^2 \tag{10.29}$$

where $\bar{P}$ is the average of cluster proportions $P_i = \frac{a_i}{M_i}$, i = 1, 2, ..., N.

Estimator of variance $V(p_{c2})$:

$$\begin{aligned} v(p_{c2}) &= \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} \left(\frac{a_i}{M_i} - p_{c2}\right)^2 \\ &= \frac{N-n}{Nn(n-1)} \left[\sum_{i=1}^{n} \left(\frac{a_i}{M_i}\right)^2 - np_{c2}^2\right] \end{aligned} \tag{10.30}$$

10.7.3 *Estimator 3*

This estimator can also be used in both the cases where M_o is known, or when it is not known. Let S_m^2 and S_{am} be defined as in (7.2) with variables the cluster size m, and a, the number of units in a cluster belonging to category A, replacing x and y respectively. Then we have the following :

Estimator of population proportion that does not depend on M_o:

$$p_{c3} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i} \tag{10.31}$$

Approximate bias of estimator p_{c3} :

$$B(p_{c3}) = \frac{N-n}{Nn\bar{M}^2}(PS_m^2 - S_{am}) \tag{10.32}$$

Approximate variance of estimator p_{c3} :

$$V(p_{c3}) = \frac{N-n}{Nn\bar{M}^2(N-1)} \sum_{i=1}^{N} (a_i - PM_i)^2 \tag{10.33}$$

Estimator of variance $V(p_{c3})$:

$$v(p_{c3}) = \frac{N-n}{Nn\bar{M}^2(n-1)} \sum_{i=1}^{n} (a_i - p_{c3}M_i)^2 \tag{10.34}$$

$\bar{M}$, if unknown, is to be replaced by $\hat{\bar{M}} = \frac{1}{n}\sum_{i=1}^{n} M_i$ in (10.34).

Example 10.7

An earlier survey conducted in a rural area, comprising a development block, showed that the proportion of infants vaccinated against polio was only .05. A vigorous campaign was then launched to make the people of this area aware of the need for vaccination against this disease. The Department of Health wanted to have an idea about the extent of impact the campaign had made. This information might be helpful in framing policies for future campaigns. To accomplish the task, 20 villages out of a population of 238 villages were selected using SRS without replacement. Children in a village, who should have been vaccinated but were not vaccinated before the campaign commenced, formed the units in the cluster. All such children in the 238 villages of the block were the target population. The data on the number of children vaccinated after the launch of campaign, are presented in table 10.6 along with other intermediate computations.

Table 10.6 Data regarding children vaccinated after the launching of campaign

Village	Target children (M_i)	Vaccinated children (a_i)	a_i/M_i
1	860	63	.07326
2	935	122	.13048
3	400	105	.26250
4	825	221	.26788
5	642	151	.23520
6	406	90	.22167
7	809	150	.18541
8	679	160	.23564
9	618	103	.16667
10	331	130	.39275
11	410	103	.25122
12	1060	198	.18679
13	576	76	.13194
14	318	113	.35535
15	845	117	.13846
16	921	212	.23018
17	308	44	.14286
18	218	82	.37615
19	880	171	.19432
20	770	130	.16883
Total	12811		4.34756

Using estimator 2, examine whether the campaign for vaccination against polio has been effective ? Also, build up confidence interval for population proportion.

Solution

The statement of the example provides that N = 238 and n = 20. Also, the calculated values for a_i/M_i are given in the last column of table 10.6. Using estimator p_{c2} in (10.27) for proportion P, we get from table 10.6,

$$p_{c2} = \frac{1}{n} \sum_{i=1}^{n} \frac{a_i}{M_i} = \frac{4.34756}{20} = .21738$$

as an estimate of the proportion of children vaccinated against polio in the target population.

Estimate of variance is calculated from (10.30). This is

$$v(p_{c2}) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} \left(\frac{a_i}{M_i} - p_{c2}\right)^2$$

$$= \frac{238-20}{(238)(20)(19)} [(.07326-.21738)^2 + (.13048-.21738)^2$$

$$+ \ldots + (.16883-.21738)^2]$$

$$= \frac{238-20}{(238)(20)(19)} [(.07326)^2 + (.13048)^2 + \ldots + (.16883)^2 - 20(.21738)^2]$$

$$= \frac{238-20}{(238)(20)(19)} (.13637)$$

$$= .000329$$

Then we compute the required confidence interval for population proportion P from

$$p_{c2} \pm 2\sqrt{v(p_{c2})}$$

$$= .21738 \pm 2\sqrt{.000329}$$

$$= .21738 \pm .03628$$

$$= .18110, .25366$$

The confidence limits above indicate that the proportion of children vaccinated in the target population of 238 villages, after the campaign was launched, is most likely to fall in the closed interval [.18110, .25366]. ■

10.8 SAMPLE SIZE REQUIRED FOR ESTIMATION OF PROPORTION

Analogous to sample size determination in case of continuous data, a preliminary sample of size n_1 is drawn. Proceeding in the same way as in section 10.6, one gets the required size for a sample of clusters to estimate the population proportion with a specified error that can be tolerated.

Sample size required to estimate proportion P with B as tolerable error :

$$n = \frac{Ns_{ci}^2}{ND + s_{ci}^2} \tag{10.35}$$

where D and s_{ci}^2 for the three estimators are defined below :

Estimator 1 :

$$s_{c1}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(\frac{a_i}{M} - p_{c1}\right)^2 \tag{10.36}$$

$$D = \frac{B^2}{4}$$

Estimator 2 :

$$s_{c2}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(\frac{a_i}{M_i} - p_{c2} \right)^2 \qquad (10.37)$$

$$D = \frac{B^2}{4}$$

Estimator 3 :

$$s_{c3}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (a_i - M_i p_{c3})^2 \qquad (10.38)$$

$$D = \frac{B^2 \bar{M}^2}{4}$$

As before, if $\bar{M}$ is unknown, it is to be estimated from the preliminary sample. Also, if $n_1 \geq n$, no additional cluster need to be selected, otherwise, $(n-n_1)$ more clusters will be selected to get the required overall sample.

Example 10.8

The data of example 10.7 had been collected from a sample of 20 villages. Assuming this sample as a preliminary sample, determine the sample size required to estimate the proportion of children vaccinated with a bound of magnitude .03 on the error of estimation.

Solution

In this problem, we have N=238 and the preliminary sample size $n_1 = 20$. Then on using (10.37) and the other intermediate computations for estimate of variance from example 10.7, we work out

$$s_{c2}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(\frac{a_i}{M_i} - p_{c2} \right)^2$$

$$= \frac{.13637}{19}$$

$$= .007177$$

Further,

$$D = \frac{B^2}{4} = \frac{(.03)^2}{4} = .000225$$

Then the sample size needed to estimate the population proportion under study with a permissible error of .03, can be obtained by using (10.35) as

$$n = \frac{N s_{c2}^2}{ND + s_{c2}^2}$$

$$= \frac{238\,(.007177)}{(238)\,(.000225) + .007177}$$

$$= 28.13$$
$$\approx 28$$

The optimum sample size, obtained above, means that if the investigator wishes to estimate the population proportion with a tolerable error of magnitude .03, he/she will have to select 28-20 = 8 more villages. ■

10.9 SELECTION OF CLUSTERS WITH UNEQUAL PROBABILITIES

In many practical situations, clusters appreciably differ in their size and the cluster total for the study variable is likely to be positively correlated with the number of units in the cluster. In such cases, it may be advantageous to select the clusters with probability proportional to the number of units in the cluster, instead of with equal probability. If the i-th cluster contains M_i units, the probability of selection for this cluster will be $P_i = M_i/M_o$, i = 1, 2, ..., N. Let a sample of n clusters be selected using probability proportional to size and WR method, the size being the number of units in the cluster. The estimator

$$\bar{y}_{c2} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i$$

of mean $\bar{Y}$, given in (10.4), becomes unbiased in this case.

Unbiased estimator of population mean $\bar{Y}$:

$$\bar{y}_{c2} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i \qquad (10.39)$$

Variance of estimator $\bar{y}_{c2}$:

$$V_p(\bar{y}_{c2}) = \frac{1}{nM_o} \sum_{i=1}^{N} M_i (\bar{Y}_i - \bar{Y})^2 \qquad (10.40)$$

Estimator of variance $V_p(\bar{y}_{c2})$:

$$v_p(\bar{y}_{c2}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{c2})^2 \qquad (10.41)$$

In certain other situations, the investigator may find some other more appropriate measures of cluster size. The selection probabilities for the clusters could then be taken proportional to this measure of size. In relations (10.42) to (10.44), we give an unbiased estimator of population mean $\bar{Y}$, and other related results, for any arbitrary set of selection probabilities $\{P_i\}$.

Unbiased estimator of population mean $\bar{Y}$:

$$\bar{y}_{c4} = \frac{1}{nM_o} \sum_{i=1}^{n} \frac{y_{i.}}{p_i} \tag{10.42}$$

Variance of estimator $\bar{y}_{c4}$:

$$V(\bar{y}_{c4}) = \frac{1}{M_o^2 n} \sum_{i=1}^{N} \left(\frac{Y_{i.}}{P_i} - Y.. \right)^2 P_i \tag{10.43}$$

Estimator of variance $V(\bar{y}_{c4})$:

$$v(\bar{y}_{c4}) = \frac{1}{M_o^2\, n(n-1)} \left(\sum_{i=1}^{n} \frac{y_{i.}^2}{p_i^2} - nM_o^2\, \bar{y}_{c4}^2 \right) \tag{10.44}$$

Estimators of population total Y.. are obtained by multiplying the estimators of population mean $\bar{Y}$ by M_o. The expressions for variance and estimators of variance for these estimators are obtained, as before, by multiplying corresponding expressions for $\bar{Y}$ by M_o^2.

Example 10.9

An agricultural university consists of N = 56 departments including 22 research stations. There are M_o= 570 fast moving vehicles (cars, jeeps, etc.) in the university. The number of vehicles in a department (M_i) vary with the strength of faculty and the nature of research work being carried out. The objective of the survey is to estimate the number of unsafe mounted tires for all the vehicles. For this purpose, a WR sample of n = 14 departments /research stations (D/RS) was selected with probability proportional to the number of vehicles in the department. A team of experts examined all the vehicles in the selected departments /research stations. The information thus collected, is given in the table below :

Table 10.7 Number of vehicles (M_i) and unsafe mounted tires ($y_{i.}$) for the selected D/RS

D/RS	M_i	$y_{i.}$	D/RS	M_i	$y_{i.}$
1	9	4	8	14	9
2	5	3	9	19	12
3	11	8	10	10	6
4	17	9	11	8	3
5	10	6	12	6	2
6	6	0	13	9	3
7	12	3	14	19	9

Estimate the total number of unsafe mounted tires being used in all the 570 vehicles, and place confidence limits on this total.

Solution

We have $N = 56$, $M_o = 570$, and $n = 14$. Through (10.39), the estimator of total is written as

$$\hat{Y}_{c2} = M_o \bar{y}_{c2} = \frac{M_o}{n} \sum_{i=1}^{n} \frac{y_{i.}}{M_i}$$

On making substitutions, one gets the estimate of total number of unsafe tires as

$$\hat{Y}_{c2} = \frac{570}{14}\left(\frac{4}{9} + \frac{3}{5} + \ldots + \frac{9}{19}\right)$$

$$= 266.31$$

$$\approx 266$$

The estimator of variance $V(\hat{Y}_{c2})$, from (10.41), will be

$$v_p(\hat{Y}_{c2}) = M_o^2 \, v_p(\bar{y}_{c2})$$

$$= \frac{M_o^2}{n(n-1)} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{c2})^2$$

$$= \frac{M_o^2}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_{i.}}{M_i} - \frac{\hat{Y}_{c2}}{M_o}\right)^2$$

$$= \frac{(570)^2}{14(14-1)}\left[\left(\frac{4}{9} - .467\right)^2 + \left(\frac{3}{5} - .467\right)^2 + \ldots + \left(\frac{9}{19} - .467\right)^2\right]$$

$$= \frac{(570)^2 \, (.492645)}{14(14-1)}$$

$$= 879.453$$

Following (2.8), we now calculate the confidence limits for the total number of unsafe tires being used in all the 570 vehicles. These limits are

$$\hat{Y}_{c2} \pm 2\sqrt{v_p(\hat{Y}_{c2})}$$

$$= 266.31 \pm 2\sqrt{879.453}$$

$$= 266.31 \pm 59.31$$

$$= 207.00, 325.62$$

$$\approx 207, 326$$

It can, therefore, be concluded that the total number of unsafe tires mounted on all the 570 vehicles will, most probably, range from 207 to 326. ■

10.10 SOME FURTHER REMARKS

10.1 For a given total number of units in a cluster sample, the sampling variance increases with the increase in cluster size (which consequently results in the decrease of the number of clusters in the sample). On the other hand, the survey cost decreases with the increase in cluster size. In surveys it is, therefore, imperative to strike a balance between the two opposing points. This problem has been considered by various workers. The details are available in Murthy (1967) and Sukhatme *et al.* (1984).

10.2 The cluster sampling, though cheaper, is generally less efficient as compared to the usual simple random sampling of population units. However, if auxiliary information in the form of values of the study variable from the recent past is available, the efficiency of cluster sampling could possibly be improved. Zarkovich and Krane (1965) have considered this aspect of cluster sampling.

10.3 The discussion in this chapter so far has been limited to nonoverlapping clusters where every population unit belongs to one and only one cluster. However, there could be situations where certain population units may fall in more than one cluster. Such clusters are known as *overlapping clusters*. The problem of estimating population mean in such cases has been considered by Tracy and Osahan (1994).

LET US DO

10.1 Define cluster sampling. In which situation is it expected to work better in relation to the usual simple random sampling ?

10.2 In what way the cluster sampling is different from the stratified sampling ? Discuss.

10.3 Give expressions for the unbiased estimator of mean and the estimator of its variance when total number of units in the population is known. From these expressions, how will you write expressions for unbiased estimator of total and the estimator of its variance ?

10.4 '*Ramayana*', the famous religious TV serial of India, was telecast for about 1 year and 9 months. Later, a private company released videocassette of this serial. The company has 30 centers throughout India, and each center has its dealers. In all, there are 250 dealers. After 6 months, an investigator wanted to estimate the total number of times a videocassette of the serial was hired after its introduction in the market. For this purpose, six centers were selected using WOR simple random sampling, and all the dealers in the sample centers were enumerated for the number of times videocassette tapes of the serial were hired from the dealers. The information, so obtained, is given in the following table.

Center	Dealers	Number of times videocassette was hired
1	7	10 19 7 4 16 21 13
2	3	8 13 14
3	6	6 9 22 5 17 8
4	4	4 2 17 13
5	9	12 9 21 24 18 7 12 19 10
6	4	6 11 12 7

Estimate the average number of times a videocassette of the serial was hired from a dealer, and place confidence limits on its population value.

10.5 What are the two estimators of mean in cluster sampling that can be used in situations when the total number of population units is not known ? Also, give expressions for variances of these estimators.

10.6 A sociologist is interested in determining the average degree of mental alertness in persons of age over 80 years in a development block consisting of 70 villages. Since the number of such individuals was not known for each village, a WOR simple random sample of 9 villages was selected. All the persons, with age over 80 years in each selected village, were interviewed and then alertness was ranked from 0 to 10. The score of 10 was given to the persons with perfect mental alertness, whereas 0 score was meant for persons who were not in control of their mental faculties. The number of persons with age over 80 years (M_i) and the scores for their mental alertness, are given below for the sample villages.

Village	M_i	Scores
1	7	4 6 1 3 7 9 2
2	11	3 5 2 9 6 8 4 10 5 3 6
3	4	5 7 9 4
4	5	6 9 8 2 5
5	8	1 5 9 8 4 3 8 10
6	6	6 8 4 2 5 7
7	4	5 7 9 1
8	7	2 4 9 7 10 1 10
9	9	3 6 2 4 9 10 1 9 5

Estimate the average mental alertness for the population of persons aged above 80 years in the development block, and place confidence limits on this average.

10.7 The Department of Education of a state has been providing fixed medical allowance at the rate of rupees 60 per head, for a quarter, to its teachers and their dependents for the last five years. With a view to examine the rationality of this policy today, when the price index has gone up about 1.5 times during the preceding five years, a simple random WOR sample of 10 schools was drawn from a total of 104 schools in a development block by the investigator. Since some of the teachers might be

on long leave, the total number of teachers available could not be known in advance. All the teachers (M_i), except those on long leave, in the sample schools were interviewed. They were requested to give per head medical expenses (in rupees), for themselves and their dependents, during the past 3 months. The results are as follows :

School	M_i	Per head medical expenses for 3 months						
1	4	50	100	120	110			
2	4	90	70	40	140			
3	5	85	33	122	60	105		
4	6	55	80	130	70	240	80	
5	4	130	70	40	120			
6	3	85	65	45				
7	4	30	75	65	115			
8	6	150	105	0	25	185	100	
9	5	100	60	130	40	125		
10	7	50	45	110	120	60	140	95

Estimate the average per head money spent as medical expenses during the past 3 months, using the estimators given in (10.4) and (10.8). Also, build up the confidence interval for the population average in each case.

10.8 From the sample observations in exercise 10.4, estimate the total number of times the videocassettes of the serial '*Ramayana*' were hired from all the centers by using the estimator given in (10.14). Also, place confidence limits on this population total.

10.9 There are 300 ponds in a development block consisting of 120 villages. The block administration is planning to use these ponds for fish farming. It is felt that the actual area of ponds is less than that appearing in revenue records. It is perhaps due to the encroachment of the pond area by the residents. A survey was, therefore, undertaken to estimate the total pond area, available at present, in the block. A WOR simple random sample of 10 villages was drawn. Area of each pond in the sample villages was accurately measured. The number of ponds (M_i) in each sample village and the area of ponds are as follows :

Village	M_i	Area (in hectares)			
1	3	2.56	.43	1.62	
2	2	1.31	2.94		
3	4	1.05	2.66	.87	1.02
4	2	2.31	1.75		
5	4	.22	.85	1.93	.74
6	2	1.36	.99		
7	2	.34	1.41		
8	3	2.07	1.61	.73	
9	4	.73	1.82	1.16	.58
10	4	.42	.81	1.3	.75

Using the estimator in (10.14), determine the total pond area in the development block, and also place confidence limits on its population value.

10.10 How will you determine the optimal number of clusters to be included in the sample, in case the estimator given in (10.14) is to be used, for estimating population total with a margin of error of magnitude B ?

10.11 Assume that the sample of 15 buses selected in example 10.2 is a preliminary sample. Using the information obtained from that sample, determine how many more buses are to be included in the sample if the variance of mean is fixed at .007?

10.12 Which of the three estimators given in (10.24), (10.27), and (10.31) will you prefer for estimation of population proportion in case the total number of units in the population is known ? State your reasons.

10.13 A survey project was undertaken to estimate the proportion of railway trains reaching late at their terminal station, in a certain zone. For this purpose, 14 stations from a total of 105 terminal stations were selected using WOR equal probability sampling. The total number of trains terminating at all these 105 stations were counted from railway time table and was found to be 773. All the trains terminating at the selected terminal stations were examined for 24 hours, for late arrivals. A train was considered to be late if it moved into the station 10 or more minutes behind schedule. The information regarding the total number of trains terminating (M_i) and the number of trains arriving late (a_i) at the sample stations is given in the table below:

Station	M_i	a_i	Station	M_i	a_i
1	34	9	8	6	2
2	4	1	9	9	2
3	10	2	10	15	3
4	16	4	11	24	5
5	8	1	12	11	2
6	28	6	13	5	0
7	15	3	14	13	3

Estimate unbiasedly, the proportion/ total number of trains arriving late at all the 105 terminal stations. Also, work out the standard error of your estimate, and place confidence limits on the population parameter being estimated.

10.14 On a canal distributary, there are 56 outlets in the last 10 km length from where the water is supplied to farmers' fields for irrigation. The exact number of farmers using water from these 56 outlets, was not known because of sale and purchase of agricultural land over time. The objective of survey was to assess whether the farmers at the end of distributary were satisfied with the release of water during paddy season, or some more water was required to be released. A WOR simple random sample of 9 outlets was selected, and the views of all the users of the

sample outlets were elicited. The response 1 indicates that the farmer was satisfied with the present supply of water, whereas 0 response indicates that additional water needs to be released. The data collected are given in the table below :

Outlets	Users	Response of users
1	4	1 0 0 1
2	7	0 1 1 1 0 1 0
3	8	1 0 0 1 1 1 1 0
4	9	1 0 1 0 1 1 1 1 0
5	6	0 1 0 1 0 0
6	4	1 1 1 0
7	7	1 0 0 1 1 1 1
8	5	0 1 1 1 1
9	9	1 0 0 1 0 1 1 1 1

Estimate proportion of farmers who are satisfied with the present supply of water, and also construct the confidence interval for it.

10.15 Taking the sample of 14 terminal stations drawn in exercise 10.13 as a preliminary sample, work out the required number of stations to be included in the sample, so that, the proportion of late arriving trains can be estimated with a margin of error .06.

10.16 Give expressions for the estimator of population mean and the estimator of its variance if the clusters are selected using varying probability WR sampling.

10.17 Suppose that the sample of 12 villages in example 10.1 was selected using PPS with replacement, the size measure being the number of farmers in the village. Using the sample data in table 10.1, estimate the parameter in question.

CHAPTER 11

Multistage Sampling

11.1 INTRODUCTION

In chapter 10, we have considered sampling procedures in which all the elements of the selected clusters are enumerated. It was seen that though cluster sampling is generally economical, but it is usually less efficient than sampling of same number of ultimate units directly from the population. This is because the former strategy restricts the spread of the sample over the population. It can, therefore, be logically expected that, for a given number of units in the sample, greater precision can be attained if (1) the units are distributed over a larger number of clusters, and (2) instead of completely enumerating all the units in each selected cluster, only a sample of units is observed. This logic gives rise to the following definition :

Definition 11.1 The procedure of sampling, which consists in first selecting the clusters and then randomly choosing a specified number of units from each selected cluster, is known as *two-stage sampling.*

The clusters that form the units of sampling at the first stage are called the *first stage units*, or *primary stage units*, and the elements within the clusters which form the units of sampling at the second stage are called *subunits*, or *secondary units,* or *second stage units*. The procedure can be generalized to three or more stages, and is then termed *multistage sampling*. As an example of four-stage sampling, we consider surveys for estimating yield of a crop in a particular state. Here, the development blocks may be considered as primary stage units, villages within blocks the second stage units, fields within villages the third stage units, and the small plots within fields, which are harvested to record yield, as the fourth stage units. It may be mentioned that multistage sampling may be the only feasible procedure in a number of practical situations, where a satisfactory sampling frame of ultimate observational units is not readily available and the cost of obtaining such a frame is considerable.

Multistage sampling is being currently used in a number of surveys. Among the early workers, Mahalanobis (1940) used this sampling procedure in estimating area under jute. The use of this procedure in agriculture and population surveys respectively has been considered by Cochran (1939) and Hansen and Hurwitz (1943). Lahiri (1954) has discussed the use of this procedure in the Indian National Sample Surveys.

Keeping the scope of the book in view, we shall only consider two-stage sampling. For the extension of the estimation procedure to more than two-stage sampling, the reader may refer to Sukhatme *et al.* (1984), Murthy (1967), and Cochran (1977).

11.2 NOTATIONS

For the two-stage estimation procedure, we shall use the following notations :

N = number of primary stage units (psu's) in the population

n = number of psu's selected in the sample

M_i = number of second stage units (ssu's) in the i-th psu

m_i = number of ssu's selected from M_i ssu's

$M_o = \sum_{i=1}^{N} M_i$ = total number of ssu's in the population

$\overline{M}$ = M_o/N = average number of ssu's per psu

Y_{ij} = value of the estimation variable y for j-th ssu of the i-th psu, $j = 1, 2, ..., M_i$; $i = 1, 2, ..., N$

y_{ij} = value of the study variable for j-th selected ssu of the i-th selected psu, $j = 1, 2, ..., m_i$; $i = 1, 2, ..., n$

$Y_{i.} = \sum_{j=1}^{M_i} Y_{ij}$ = total of y-values for the i-th psu

$Y.. = \sum_{i=1}^{N} Y_{i.}$ = population total of y-values

$\overline{Y}_i = \frac{Y_{i.}}{M_i}$ = population mean for i-th psu

$\overline{Y}_N = \frac{1}{N} \sum_{i=1}^{N} \overline{Y}_i$ = population mean of psu means

$\overline{Y} = \frac{1}{M_o} \sum_{i=1}^{N} \overline{Y}_i M_i$ = population mean per ssu for the study variable

$y_{i.} = \sum_{j=1}^{m_i} y_{ij}$ = sample total for the i-th psu

$y.. = \sum_{i=1}^{n} y_{i.}$ = total of y-values for the whole sample

$\overline{y}_i = \frac{y_{i.}}{m_i}$ = sample mean for i-th psu

11.3 ESTIMATION OF MEAN/ TOTAL IN TWO-STAGE SAMPLING USING SRSWOR AT BOTH THE STAGES

We consider two-stage sampling where the first stage units are of unequal size, and the units are selected using equal probabilities WOR method at both the stages. To select the sample, the investigator must have a frame listing all the N psu's in the population. A WOR simple random sample of n psu's is drawn using procedures described in

chapter 3. Then, the frames that list all the M_i second stage units in i-th selected psu (i =1, 2 ,..., n) are obtained. Finally, a WOR simple random sample of m_i units is drawn from the i-th selected psu, containing M_i second stage units, for i = 1, 2,.., n. Several estimators of mean and total are available for this sampling procedure. However, we shall consider only three of these estimators as they are used quite frequently.

11.3.1 *Estimator 1*

Let us define

$$\left.\begin{aligned} S_{1b}^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i\,\bar{Y}_i}{\bar{M}}-\bar{Y}\right)^2 \\ &= \frac{1}{\bar{M}^2\,(N-1)}\left(\sum_{i=1}^{N}Y_{i.}^2-\frac{Y_{..}^{\,2}}{N}\right) \end{aligned}\right] \qquad (11.1)$$

$$\left.\begin{aligned} S_i^2 &= \frac{1}{M_i-1}\sum_{j=1}^{M_i}(Y_{ij}-\bar{Y}_i)^2 \\ &= \frac{1}{M_i-1}\left(\sum_{j=1}^{M_i}Y_{ij}^2-M_i\,\bar{Y}_i^2\right) \end{aligned}\right] \qquad (11.2)$$

$$\left.\begin{aligned} s_{1b}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i\bar{y}_i}{\bar{M}}-\bar{y}_{ml}\right)^2 \\ &= \frac{1}{\bar{M}^2\,(n-1)}\left[\sum_{i=1}^{n}(M_i\,\bar{y}_i)^2-\frac{1}{n}\left(\sum_{i=1}^{n}M_i\,\bar{y}_i\right)^2\right] \end{aligned}\right] \qquad (11.3)$$

$$\left.\begin{aligned} s_i^2 &= \frac{1}{m_i-1}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2 \\ &= \frac{1}{m_i-1}\left(\sum_{j=1}^{m_i}y_{ij}^2-m_i\bar{y}_i^2\right) \end{aligned}\right] \qquad (11.4)$$

where $\bar{y}_{ml}$ is given in (11.5). Then we have expressions (11.5) to (11.7).

Unbiased estimator of population mean $\bar{Y}$:

$$\bar{y}_{ml} = \frac{N}{nM_o}\sum_{i=1}^{n}M_i\,\bar{y}_i \qquad (11.5)$$

Variance of the estimator $\bar{y}_{ml}$:

$$V(\bar{y}_{ml}) = \left(\frac{1}{n}-\frac{1}{N}\right)S_{1b}^2+\frac{1}{nN}\sum_{i=1}^{N}\frac{M_i^2}{\bar{M}^2}\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^2 \qquad (11.6)$$

Estimator of variance $V(\bar{y}_{ml})$:

$$v(\bar{y}_{ml}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{1b}^2 + \frac{1}{nN} \sum_{i=1}^{n} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 \qquad (11.7)$$

where S_{1b}^2, S_i^2, s_{1b}^2, and s_i^2 are defined in (11.1) through (11.4).

Example 11.1

The co-operative societies in an Indian state, provide loans to farmers in terms of cash and fertilizer within the sanctioned limit, which depends on the share of the individual in the co-operative society. The society declares an individual defaulter, if he/she does not repay the loan within the specified time limit. An investigator is interested in estimating the average amount of loan, per society, standing against the defaulters. The total number of co-operative societies in the state is 10126. However, the list of all the societies is not available at the state headquarter but the same is available at development block level. Therefore, it seems appropriate to use two-stage sampling for selecting a sample of societies. Keeping in view the budget and time constraints, it was decided to select 12 blocks from the total of 117 blocks and approximately 10 percent of the societies from each of the sample blocks. The information obtained from the selected societies is given in table 11.1

Table 11.1 Dues (in '000 rupees) standing against the defaulters

Block	M_i	m_i	Amount due from defaulters						Total
1	60	6	12.5	36.4	26.0	55.6	58.1	40.8	229.4
2	102	10	57.4	16.8	20.3	70.1	34.6	22.6	346.3
			44.9	28.4	17.5	33.7			
3	48	5	12.9	41.6	34.7	30.8	61.1		181.1
4	113	11	28.7	82.4	37.3	41.9	24.7	36.6	494.9
			39.3	49.6	26.0	76.8	51.6		
5	92	9	44.8	42.9	51.7	28.8	36.4	40.1	431.5
			61.6	47.8	77.4				
6	57	6	31.6	24.8	69.9	44.9	59.7	38.6	269.5
7	82	8	49.6	36.9	27.3	63.6	73.0	44.9	443.6
			87.1	61.2					
8	96	10	53.7	34.9	41.5	43.4	56.6	28.9	423.0
			23.4	32.8	60.2	47.6			
9	53	5	41.7	54.9	33.9	27.9	46.3		204.7
10	71	7	24.4	38.9	47.8	45.0	32.6	66.5	313.5
			58.3						
11	77	8	42.9	37.3	30.8	51.9	60.1	34.6	324.3
			28.4	38.3					
12	56	6	44.7	34.9	61.7	74.6	37.4	49.2	302.5

Estimate the average amount, per society, standing against defaulters, and also compute the confidence interval for it.

Solution

The statement of the problem shows that N = 117, n = 12, and $M_0 = 10126$. Hence,

$$\overline{M} = \frac{10126}{117} = 86.55$$

In order to illustrate the various steps involved in the calculations, we prepare table 11.2. The mean $\bar{y}_i$ and the sample mean square s_i^2 for the societies included in the sample, for each selected block, are calculated. These are given along with other computations in table 11.2.

Table 11.2 Calculations for various terms involved in the estimation of population mean

Block	M_i	m_i	$\bar{y}_i = \frac{y_i}{m_i}$	s_i^2	$M_i^2 (\frac{1}{m_i} - \frac{1}{M_i}) s_i^2$	$M_i\bar{y}_i$
1	60	6	38.23	303.62	163954.80	2293.80
2	102	10	34.63	320.35	300616.44	3532.26
3	48	5	36.22	305.87	126263.13	1738.56
4	113	11	44.99	368.54	386162.91	5083.87
5	92	9	47.94	208.11	176569.77	4410.48
6	57	6	44.92	292.93	141924.58	2560.44
7	82	8	55.45	384.47	291620.49	4546.90
8	96	10	42.30	151.84	125359.10	4060.80
9	53	5	40.94	110.95	56451.36	2169.82
10	71	7	44.79	210.33	136534.21	3180.09
11	77	8	40.54	115.75	76872.47	3121.58
12	56	6	50.42	231.30	107940.00	2823.52
Total	907				2090269.26	39522.12

Since M_0 is available, we use (11.5) for obtaining unbiased estimator of population mean. Thus, the estimate of average dues, per society, standing against defaulters is given by

$$\bar{y}_{ml} = \frac{N}{nM_0} \sum_{i=1}^{n} M_i\bar{y}_i$$

$$= \frac{117}{(12)(10126)} [(60)(38.23) + (102)(34.63) + \ldots + (56)(50.42)]$$

$$= \frac{(117)(39522.12)}{(12)(10126)}$$

$$= 38.05$$

The estimator of variance $V(\bar{y}_{ml})$ is given by (11.7). Hence,

$$v(\bar{y}_{ml}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{1b}^2 + \frac{1}{nN} \sum_{i=1}^{n} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2$$

The expression for s_{1b}^2 is given in (11.3). We calculate it by using the last column of table 11.2 as

$$s_{1b}^2 = \frac{1}{(86.55)^2\,(12-1)}[(2293.80)^2 + (3532.26)^2 + \ldots + (2823.52)^2 - \frac{1}{12}(39522.12)^2]$$
$$= 147.46$$

On making substitutions from table 11.2, one gets

$$v(\bar{y}_{ml}) = \left(\frac{1}{12} - \frac{1}{117}\right)(147.46) + \frac{2090269.26}{(12)\,(117)\,(86.55)^2}$$
$$= 11.03 + .20$$
$$= 11.23$$

The confidence limits for the population mean are given by

$$\bar{y}_{ml} \pm 2\sqrt{v(\bar{y}_{ml})}$$
$$= 38.05 \pm 2\sqrt{11.23}$$
$$= 38.05 \pm 6.70$$
$$= 31.35,\ 44.75$$

These limits yield the confidence interval as [31.35, 44.75]. It means that the investigator can reasonably believe that if whole of the population is enumerated, the average dues, per society, standing against the defaulters are likely to be in the range of 31.35 to 44.75 thousand rupees. ■

The estimator $\bar{y}_{ml}$ in (11.5), though unbiased, uses the knowledge of M_o. In practice, the situations may arise where the value of M_o (or equivalently $\bar{M}$) is not available. For such cases, the estimators that do not depend on M_o are needed. We shall now consider two such estimators of mean $\bar{Y}$.

11.3.2 *Estimator 2*

Again, let

$$\left.\begin{aligned} S_{2b}^2 &= \frac{1}{N-1} \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y}_N)^2 \\ &= \frac{1}{N-1} \left(\sum_{i=1}^{N} \bar{Y}_i^2 - N\bar{Y}_N^2\right) \end{aligned}\right] \qquad (11.8)$$

$$\left.\begin{aligned} s_{2b}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{y}_{m2})^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\bar{y}_i^2 - n\bar{y}_{m2}^2\right) \end{aligned}\right] \qquad (11.9)$$

where $\bar{y}_{m2}$ is obtained from (11.10). The expressions for bias, variance, and estimator of variance for this estimator are given in relations (11.11) through (11.13).

Estimator of population mean which does not depend on M_o :

$$\bar{y}_{m2} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i \qquad (11.10)$$

Bias of the estimator $\bar{y}_{m2}$:

$$B(\bar{y}_{m2}) = -\frac{1}{N\bar{M}}\sum_{i=1}^{N}(M_i - \bar{M})\bar{Y}_i \qquad (11.11)$$

Variance of the estimator $\bar{y}_{m2}$:

$$V(\bar{y}_{m2}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_{2b}^2 + \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_i^2 \qquad (11.12)$$

Estimator of variance $V(\bar{y}_{m2})$:

$$v(\bar{y}_{m2}) = \left(\frac{1}{n} - \frac{1}{N}\right)s_{2b}^2 + \frac{1}{nN}\sum_{i=1}^{n}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_i^2 \qquad (11.13)$$

The expressions for S_{2b}^2, S_i^2, s_{2b}^2, and s_i^2 are given in (11.8), (11.2), (11.9), and (11.4) respectively.

It can be seen that the first term in (11.6) depends on the variation between psu totals. This component is larger than the corresponding component in (11.12), provided the correlation between psu size and the psu mean is positive, and the bias is small. The second term of (11.6) is also likely to be more than the second term of (11.12) as there is expected to be a positive correlation between M_i and S_i^2. Because of these considerations, the estimator $\bar{y}_{m2}$ should be preferred over the estimator $\bar{y}_{m1}$ unless the bias in $\bar{y}_{m2}$ is serious. We now consider another estimator of mean which does not depend on the knowledge of M_o. Thus, it can be used irrespective of the availability of information on M_o.

11.3.3 *Estimator 3*

We further define

$$\left.\begin{aligned} S_m^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(M_i - \bar{M})^2 \\ &= \frac{1}{N-1}\left(\sum_{i=1}^{N}M_i^2 - N\bar{M}^2\right) \end{aligned}\right] \qquad (11.14)$$

$$S_{my} = \frac{1}{N-1}\sum_{i=1}^{N}(M_i - \bar{M})(M_i\bar{Y}_i - \bar{Y}\bar{M}) = \frac{1}{N-1}\left(\sum_{i=1}^{N} M_i Y_{i.} - N\bar{Y}\bar{M}^2\right) \quad (11.15)$$

$$S_{3b}^2 = \frac{1}{\bar{M}^2(N-1)}\sum_{i=1}^{N} M_i^2(\bar{Y}_i - \bar{Y})^2 \quad (11.16)$$

$$s_{3b}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\frac{M_i^2}{\bar{M}^2}(\bar{y}_i - \bar{y}_{m3})^2 \quad (11.17)$$

where $\bar{y}_{m3}$ has been defined in (11.18).

Estimator of population mean which does not depend on M_o :

$$\bar{y}_{m3} = \frac{\sum_{i=1}^{n} M_i\bar{y}_i}{\sum_{i=1}^{n} M_i} \quad (11.18)$$

Approximate bias of the estimator $\bar{y}_{m3}$:

$$B(\bar{y}_{m3}) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{\bar{M}^2}(\bar{Y}S_m^2 - S_{my}) \quad (11.19)$$

Approximate variance of the estimator $\bar{y}_{m3}$:

$$V(\bar{y}_{m3}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_{3b}^2 + \frac{1}{\bar{M}^2 nN}\sum_{i=1}^{N} M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_i^2 \quad (11.20)$$

Estimator of variance $V(\bar{y}_{m3})$:

$$v(\bar{y}_{m3}) = \left(\frac{1}{n} - \frac{1}{N}\right)s_{3b}^2 + \frac{1}{\bar{M}^2 nN}\sum_{i=1}^{n} M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_i^2 \quad (11.21)$$

where S_m^2, S_{my}, S_{3b}^2, and s_{3b}^2 are defined in (11.14) to (11.17), whereas S_i^2 and s_i^2 are defined in (11.2) and (11.4) respectively.

The estimator $\bar{y}_{m3}$ is likely to be more efficient than $\bar{y}_{m1}$ and $\bar{y}_{m2}$, provided n is large and correlation coefficient between $M_i\bar{Y}_i$ and M_i is positive and greater than $CV(M_i)/2\,CV(M_i\bar{Y}_i)$.

Example 11.2

An orchard owner is to sell a truckload of oranges. The oranges are packed into 140 cartons. The number of oranges per carton may vary, and the total number of oranges in the truck is also not known exactly. Before striking the deal, the buyer thinks it wise to have an idea regarding the quantity of juice in the oranges. To do this, the buyer selects 10 cartons at random and then selects approximately 5 percent of the oranges from each

selected carton. The information in respect of total number of oranges, and the juice obtained from the selected oranges, is given in table 11.3.

Table 11.3 Quantity of juice (in ml) for sample oranges

Carton	M_i	m_i	Juice (in ml)							Total
1	105	5	90	103	76	84	89			442
2	120	6	107	80	72	110	70	84		523
3	95	5	104	93	83	76	91			447
4	132	7	86	93	101	81	77	99	109	646
5	111	6	91	97	85	110	101	80		564
6	117	6	88	84	99	106	92	78		547
7	86	4	101	100	91	113				405
8	122	6	109	78	89	91	90	98		555
9	130	7	114	101	96	108	80	103	108	710
10	119	6	87	94	90	109	102	84		566

Estimate the average quantity of juice per orange, and also work out the probable range wherein the population mean would have fallen if all the oranges in the truck were observed.

Solution

In this problem, we are given that N = 140 and n = 10. Since total number of oranges M_0 (or equivalently $\overline{M}$) is not known, we can use either of the estimators given in (11.10) and (11.18). In this illustration, we proceed with estimator 2 given in (11.10). As in example 11.1, we prepare table 11.4 which provides details for the various steps involved in the solution.

Table 11.4 Calculations required for estimating population mean

Carton	M_i	m_i	$y_{i.}$	$\bar{y}_i = \frac{y_{i.}}{m_i}$	s_i^2	$(\frac{1}{m_i} - \frac{1}{M_i}) s_i^2$
1	105	5	442	88.40	97.30	18.53
2	120	6	523	87.17	300.17	47.53
3	95	5	447	89.40	112.30	21.28
4	132	7	646	92.29	133.57	18.07
5	111	6	564	94.00	120.00	18.92
6	117	6	547	91.17	103.37	16.34
7	86	4	405	101.25	81.58	19.45
8	122	6	555	92.50	106.70	16.91
9	130	7	710	101.43	122.62	16.57
10	119	6	566	94.33	90.67	14.35
Total	1137			931.94		207.95

From (11.10) and the table 11.4, we have the estimator of average quantity of juice per orange as

$$\bar{y}_{m2} = \frac{931.94}{10} = 93.19$$

Using (11.9), let us first compute s_{2b}^2. Thus we have

$$\begin{aligned} s_{2b}^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{m2})^2 \\ &= \frac{1}{10-1} [(88.40-93.19)^2 + (87.17-93.19)^2 + \ldots + (94.33-93.19)^2] \\ &= \frac{1}{10-1} [(88.40)^2 + (87.17)^2 + \ldots + (94.33)^2 - 10(93.19)^2] \\ &= 23.75 \end{aligned}$$

Making use of the calculated value of s_{2b}^2 and column (7) of table 11.4, we work out the estimate of variance from (11.13). Therefore,

$$\begin{aligned} v(\bar{y}_{m2}) &= (\frac{1}{10} - \frac{1}{140})(23.75) + \frac{207.95}{(10)(140)} \\ &= 2.354 \end{aligned}$$

As required in the statement of the problem, the lower and upper limits of the range, wherein the population mean is expected to fall, are obtained as

$$\begin{aligned} &\bar{y}_{m2} \pm 2\sqrt{v(\bar{y}_{m2})} \\ &= 93.19 \pm 2\sqrt{2.354} \\ &= 93.19 \pm 3.07 \\ &= 90.12, 96.26 \end{aligned}$$

It can, therefore, be said that the actual population average will be covered by the interval [90.12, 96.26] ml, with probability approximately .95. ■

Example 11.3

Using data of example 11.2, estimate the parameter in question through the alternative estimator $\bar{y}_{m3}$ defined in (11.18).

Solution

For working out the required estimate, some of the values computed in table 11.4 will be used. The estimator $\bar{y}_{m3}$ of the population mean, given in (11.18), is

$$\bar{y}_{m3} = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}$$

On making use of columns (2) and (5) of table 11.4, we get

$$\bar{y}_{m3} = \frac{(105)(88.40) + (120)(87.17) + \ldots + (119)(94.33)}{(105+120+\ldots+119)}$$

$$= \frac{105922.24}{1137}$$

$$= 93.16$$

Thus, on using the estimator 3, the estimate of juice per orange is obtained as 93.16 ml.

Since $\bar{M}$ is not available, we shall use its sample estimate $\hat{\bar{M}}$ wherever necessary. For this we have

$$\hat{\bar{M}} = \frac{1}{n} \sum_{i=1}^{n} M_i = \frac{1137}{10} = 113.7$$

The estimator of variance from (11.21) is

$$v(\bar{y}_{m3}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{3b}^2 + \frac{1}{nN} \sum_{i=1}^{n} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2$$

For obtaining the value of $v(\bar{y}_{m3})$, we first compute the two terms involved in it. For working out the first term, we make use of columns (2) and (5) of table 11.4. After replacing $\bar{M}$ by $\hat{\bar{M}}$, the first term is

$$\frac{1}{\hat{\bar{M}}^2 (n-1)} \sum_{i=1}^{n} M_i^2 (\bar{y}_i - \bar{y}_{m3})^2 = \frac{1}{(113.7)^2 (10-1)} [(105)^2 (88.40-93.16)^2 + (120)^2$$

$$(87.17-93.16)^2 + \ldots + (119)^2 (94.33-93.16)^2]$$

$$= 22.655$$

The second term in $v(\bar{y}_{m3})$, after replacing $\bar{M}$ by $\hat{\bar{M}}$, is calculated by making use of column (7) of table 11.4. Therefore,

$$\sum_{i=1}^{n} \frac{M_i^2}{\hat{\bar{M}}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 = \frac{1}{(113.7)^2} [(105)^2 (18.53) + (120)^2 (47.53)$$

$$+ \ldots + (119)^2 (14.35)]$$

$$= 211.268$$

On substituting the values calculated above in the expression for $v(\bar{y}_{m3})$, one gets

$$v(\bar{y}_{m3}) = \left(\frac{1}{10} - \frac{1}{140}\right)(22.655) + \frac{211.268}{(10)(140)}$$

$$= 2.104 + .151$$

$$= 2.255$$

The confidence limits for the average quantity of juice per orange, for whole of the target population, are given by

$$\bar{y}_{m3} \pm 2\sqrt{v(\bar{y}_{m3})}$$

$$= 93.16 \pm 2\sqrt{2.255}$$

$$= 93.16 \pm 3.00$$

$$= 90.16, 96.16$$

One can, therefore, be reasonably sure that the average quantity of juice per orange for the whole lot of oranges in the truck is likely to be in the range of 90.16 to 96.16 ml. ■

Estimation of population total Y is a straightforward exercise from the estimators presented for mean. Estimators of total can be obtained by multiplying the estimators of mean given in (11.5), (11.10), and (11.18) by M_o.

Unbiased estimator of population total that does not depend on M_o :

$$\hat{Y}_{m1} = M_o \bar{y}_{m1} = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{y}_i \qquad (11.22)$$

Estimators of population total using M_o :

$$\hat{Y}_{m2} = M_o \bar{y}_{m2} = \frac{M_o}{n} \sum_{i=1}^{n} \bar{y}_i \qquad (11.23)$$

$$\hat{Y}_{m3} = M_o \bar{y}_{m3} = \frac{M_o \sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i} \qquad (11.24)$$

The expressions for bias in the estimators $\hat{Y}_{m2}$ and $\hat{Y}_{m3}$ can be obtained by multiplying the corresponding expressions for $\bar{y}_{m2}$ and $\bar{y}_{m3}$ respectively by M_o. Similarly, the expressions for variances and variance estimators can be arrived at by multiplying the corresponding expressions for mean by M_o^2.

Example 11.4

Assume that in example 11.1, the total number of societies in the state is not known. Using the data of that example, estimate the total amount standing against defaulters of co-operative societies in the state. Also, obtain the standard error of the estimate, and place confidence limits on the actual total amount.

Solution

Here M_o is not available. The estimator $\hat{Y}_{ml}$ given in (11.22) will, therefore, be used to estimate the population total. Thus

$$\hat{Y}_{ml} = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{y}_i$$

Making use of columns (2) and (4) of table 11.2, one gets

$$\hat{Y}_{ml} = \frac{117}{12} [(60)(38.23) + (102)(34.63) + \ldots + (56)(50.42)]$$
$$= 385340.67$$

as an estimate of the total amount standing against all the defaulters in the state.

The estimator of the variance $V(\hat{Y}_{ml})$ can be written, from (11.7), as

$$v(\hat{Y}_{ml}) = M_o^2 \, v(\bar{y}_{ml}) = N^2 \bar{M}^2 \left(\frac{1}{n} - \frac{1}{N}\right) s_{1b}^2 + \frac{N}{n} \sum_{i=1}^{n} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 \quad (11.25)$$

Since M_o, or equivalently $\bar{M}$, is unknown, we instead use its sample estimate $\hat{\bar{M}}$, where

$$\hat{\bar{M}} = \frac{1}{n} \sum_{i=1}^{n} M_i$$
$$= \frac{1}{12} (60 + 102 + \ldots + 56)$$
$$= \frac{907}{12}$$
$$= 75.58$$

For obtaining the value of s_{1b}^2, we shall now use $\hat{\bar{M}}$ in place of $\bar{M}$. Therefore, s_{1b}^2 is computed as

$$s_{1b}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{M_i \bar{y}_i}{\hat{\bar{M}}} - \bar{y}_{ml} \right)^2$$

where from example 11.1, $\bar{y}_{ml} = 38.05$. Thus on using the figures from table 11.2, we get

$$s_{1b}^2 = \frac{1}{12-1} \left[\left(\frac{(60)(38.23)}{75.58} - 38.05 \right)^2 + \left(\frac{(102)(34.63)}{75.58} - 38.05 \right)^2 \right.$$
$$\left. + \ldots + \left(\frac{(56)(50.42)}{75.58} - 38.05 \right)^2 \right]$$
$$= \frac{2493.55}{11}$$
$$= 226.69$$

Also, from table 11.2

$$\sum_{i=1}^{n} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 = 2090269.26$$

On using $\hat{\bar{M}} = 75.58$ in place of $\bar{M}$ in (11.25), and making other substitutions, one obtains

$$\begin{aligned} v(\hat{Y}_{ml}) &= (117)^2 (75.58)^2 \left(\frac{1}{12} - \frac{1}{117}\right)(226.69) + \frac{(117)(2090269.26)}{12} \\ &= 1325684115 + 20380125 \\ &= 1346064240 \end{aligned}$$

The standard error of the estimate will, therefore, be

$$\begin{aligned} se(\hat{Y}_{ml}) &= \sqrt{1346064240} \\ &= 36688.75 \end{aligned}$$

The interval in which the total amount standing against defaulters all over the state would probably fall, is given by

$$\begin{aligned} &\hat{Y}_{ml} \pm 2\, se\,(\hat{Y}_{ml}) \\ &= 385340.67 \pm 73377.50 \\ &= 311963.17,\ 458718.17 \end{aligned} \qquad (11.25)$$

Thus the total amount due is likely to range from 311963.17 to 458718.17 thousand rupees. ■

11.4 ESTIMATION OF PROPORTION

As mentioned earlier, the estimators for population proportion can be obtained from those for population mean by allowing the study variable y_{ij} to take values 1, or 0, depending on whether the j-th unit in the i-th psu falls into the category of interest, or not. Let p_i be the proportion of sampled ssu's from i-th psu that fall into specified category. Again, we consider three estimators of population proportion corresponding to the estimators of mean in (11.5), (11.10), and (11.18).

11.4.1 *Estimator 1*

We find that for a variable y taking 0 and 1 values

$$\begin{aligned} S_{1b}^2 &= \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{M_i P_i}{\bar{M}} - P\right)^2 \\ &= \frac{1}{\bar{M}^2 (N-1)} \left[\sum_{i=1}^{N} (M_i P_i)^2 - \frac{1}{N}\left(\sum_{i=1}^{N} M_i P_i\right)^2\right] \end{aligned} \qquad (11.26)$$

$$s_{1b}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i p_i}{\bar{M}} - p_{m1}\right)^2 = \frac{1}{\bar{M}^2(n-1)}\left[\sum_{i=1}^{n}(M_i p_i)^2 - \frac{1}{n}\left(\sum_{i=1}^{n} M_i p_i\right)^2\right] \quad (11.27)$$

where P_i and P are the proportions of units falling in the category under consideration in the i-th psu, and in the whole population respectively. Also, p_i and p are their sample analogs. The value of p_{m1} is computed from (11.28).

Unbiased estimator of population proportion when M_o is known :

$$p_{m1} = \frac{N}{nM_o}\sum_{i=1}^{n} M_i p_i \quad (11.28)$$

Variance of the estimator p_{m1}:

$$V(p_{m1}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{1b}^2 + \frac{1}{nN}\sum_{i=1}^{N}\frac{M_i^2}{\bar{M}^2}\left(\frac{M_i - m_i}{M_i - 1}\right)\frac{P_i Q_i}{m_i} \quad (11.29)$$

Estimator of variance $V(p_{m1})$:

$$v(p_{m1}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{1b}^2 + \frac{1}{nN\bar{M}^2}\sum_{i=1}^{n}\frac{M_i(M_i - m_i)\, p_i\, q_i}{m_i - 1} \quad (11.30)$$

where $Q_i = 1-P_i$ and $q_i = 1-p_i$. Also, the terms S_{1b}^2 and s_{1b}^2 appearing in (11.29) and (11.30) are defined in (11.26) and (11.27) respectively.

Example 11.5

A state government has 4032 harvester combines. These were allocated 5 years ago to 96 centers in the state. These centers look after the operation of combines at their disposal. Exact number of combines at any point of time, with each center, is not known as the combines are transferred from one center to the other depending on the pressure of work. The state needs to estimate the proportion of the five year old combines that are operating at loss during the current paddy harvesting season, more than 80 percent of which is already over. For this purpose, a WOR random sample of 12 centers was drawn. For each center in the sample, about 20 percent of combines were selected and the number of combines operating at loss (COL) determined. The information thus collected is presented in table 11.5.

Table 11.5 Data regarding combines, and certain other computations

Center	M_i	m_i	COL	p_i	$M_i p_i$	$\frac{M_i(M_i - m_i)\, p_i\, q_i}{m_i - 1}$
1	50	10	2	.2000	10.00	35.56
2	35	7	1	.1429	5.00	20.00
3	42	8	3	.3750	15.75	47.81
4	46	9	3	.3333	15.33	47.28
5	37	7	2	.2857	10.57	37.75
6	55	11	4	.3636	20.00	56.00
7	40	8	1	.1250	5.00	20.00
8	35	7	2	.2857	10.00	33.33
9	47	9	2	.2222	10.44	38.58
10	31	6	0	0	0	0
11	44	9	4	.4444	19.55	47.53
12	48	10	3	.3000	14.40	42.56
Total	510				136.04	426.40

Estimate the proportion of combines operating at loss, and build up the confidence interval for this proportion in the state.

Solution

The statement of the example provides that N=96, $M_o = 4032$, and n = 12. Since M_o is known, we use estimator p_{ml} defined in (11.28). This is

$$p_{ml} = \frac{N}{nM_o} \sum_{i=1}^{n} M_i\, p_i$$

Making use of total of column (6) of table 11.5, one gets the estimate of proportion of the combines operating at loss, as

$$p_{ml} = \frac{(96)\,(136.04)}{(12)\,(4032)} = .2699$$

The estimate of variance $V(p_{ml})$ is obtained by using (11.30). The expression for the variance estimator is

$$v(p_{ml}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{1b}^2 + \frac{1}{nN\overline{M}^2} \sum_{i=1}^{n} \frac{M_i(M_i - m_i)\, p_i\, q_i}{m_i - 1}$$

We then first compute

$$s_{1b}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{M_i p_i}{\overline{M}} - p_{ml}\right)^2$$

where

$$\overline{M} = \frac{4032}{96} = 42$$

Using figures from column (6) of table 11.5, we get

$$s_{1b}^2 = \frac{1}{12-1}\left[\left(\frac{10.00}{42} - .2699\right)^2 + \left(\frac{5.00}{42} - .2699\right)^2 + \ldots + \left(\frac{14.40}{42} - .2699\right)^2\right]$$

$$= \frac{.2274}{11}$$

$$= .02067$$

The value of the summation term in the second component of $v(p_{m1})$, is the total of column (7) of table 11.5. Thus we get

$$v(p_{m1}) = \left(\frac{1}{12} - \frac{1}{96}\right)(.02067) + \frac{426.40}{(12)(96)(42)^2}$$

$$= .001717$$

The required confidence interval for the population proportion is obtained from the limits

$$p_{m1} \pm 2\sqrt{v(p_{m1})}$$

$$= .2699 \pm 2\sqrt{.001717}$$

$$= .2699 \pm .0829$$

$$= .1870, .3528$$

The above confidence limits indicate with approximate probability as .95, that 18.70 to 35.28 percent combines in the population of 4032 combines are incurring loss to the state. ■

For the situations where M_o is unknown, we present estimators analogous to $\bar{y}_{m2}$ and $\bar{y}_{m3}$ given in (11.10) and (11.18) respectively.

11.4.2 *Estimator 2*

For a variable taking 0 and 1 values

$$S_{2b}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(P_i - \frac{1}{N}\sum_{i=1}^{N}P_i\right)^2$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N}P_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N}P_i\right)^2\right] \qquad (11.31)$$

$$\left.\begin{aligned} s_{2b}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(p_i - p_{m2})^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} p_i^2 - np_{m2}^2\right) \end{aligned}\right] \qquad (11.32)$$

where p_{m2} is defined in (11.33) below. Then, we have the expressions (11.33) through (11.36).

Estimator of population proportion which does not depend on M_o :

$$p_{m2} = \frac{1}{n}\sum_{i=1}^{n} p_i \qquad (11.33)$$

Bias of the estimator p_{m2} :

$$B(p_{m2}) = -\frac{1}{N\overline{M}}\sum_{i=1}^{N}(M_i - \overline{M})P_i \qquad (11.34)$$

Variance of the estimator p_{m2} :

$$V(p_{m2}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_{2b}^2 + \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{M_i - m_i}{M_i - 1}\right)\frac{P_i Q_i}{m_i} \qquad (11.35)$$

Estimator of variance $V(p_{m2})$:

$$v(p_{m2}) = \left(\frac{1}{n} - \frac{1}{N}\right)s_{2b}^2 + \frac{1}{nN}\sum_{i=1}^{n}\left(\frac{M_i - m_i}{M_i}\right)\frac{p_i q_i}{m_i - 1} \qquad (11.36)$$

The terms S_{2b}^2 and s_{2b}^2 involved in (11.35) and (11.36) above are defined in (11.31) and (11.32) respectively.

11.4.3 *Estimator 3*

For the kind of variable under consideration,

$$\left.\begin{aligned} S_{my} &= \frac{1}{N-1}\sum_{i=1}^{N}(M_i - \overline{M})(M_iP_i - \overline{M}P) \\ &= \frac{1}{N-1}\left(\sum_{i=1}^{N} M_i^2 P_i - N\overline{M}^2P\right) \end{aligned}\right] \qquad (11.37)$$

$$S_{3b}^2 = \frac{1}{\overline{M}^2(N-1)}\sum_{i=1}^{N} M_i^2(P_i - P)^2 \qquad (11.38)$$

$$s_{3b}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\frac{M_i^2}{\overline{M}^2}(p_i - p_{m3})^2 \qquad (11.39)$$

with p_{m3} defined in (11.40). We thus have :

Estimator of population proportion which does not depend on M_o :

$$p_{m3} = \frac{\sum_{i=1}^{n} M_i p_i}{\sum_{i=1}^{n} M_i} \tag{11.40}$$

Approximate bias of the estimator p_{m3} :

$$B(p_{m3}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{\overline{M}^2} (PS_m^2 - S_{my}) \tag{11.41}$$

Approximate variance of the estimator p_{m3} :

$$V(p_{m3}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{3b}^2 + \frac{1}{nN\overline{M}^2} \sum_{i=1}^{N} \left(\frac{M_i^2 (M_i - m_i)}{M_i - 1}\right) \frac{P_i Q_i}{m_i} \tag{11.42}$$

Estimator of variance $V(p_{m3})$:

$$v(p_{m3}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{3b}^2 + \frac{1}{nN\overline{M}^2} \sum_{i=1}^{n} \left(\frac{M_i^2 (M_i - m_i)}{M_i}\right) \frac{p_i q_i}{m_i - 1} \tag{11.43}$$

The terms S_m^2, S_{my}, S_{3b}^2, and s_{3b}^2 have already been defined in (11.14), (11.37), (11.38), and (11.39) respectively. When $\overline{M}$ is not known, it has to be replaced by $\hat{\overline{M}}$ in (11.43).

Example 11.6

Do only stray dog bites cause the rabies ? The answer to this question is required to frame a policy for giving dogs antirabies shots. A survey was undertaken for this purpose in an Indian state. The state has 70 hospitals where dog bite cases are treated. The addresses of persons, treated for dog bite, were available in these hospitals. However, the category of the dogs - stray or pet- who had bitten the patient, was not mentioned in the records. Keeping the resource constraints in view, a WOR simple random sample of 11 hospitals was drawn. About 5 percent of the treated persons were sampled using SRS without replacement. The data so collected, are presented in table 11.6. The observation is recorded as 1 if the rabies was caused by pet dog bite, 0 otherwise.

Table 11.6 Sample observations regarding type of dog causing rabies

Hospital	M_i	m_i	Observations													Total	p_i
1	140	7	0	1	1	0	0	0	1							3	.4286
2	203	10	1	0	0	0	0	0	1	1	0	0				3	.3000
3	91	5	0	1	0	0	1									2	.4000
4	176	9	1	0	0	1	0	1	0	0	0					3	.3333
5	121	6	0	1	0	0	0	0								1	.1667
6	263	13	1	0	1	1	0	0	0	1	0	0	0	1	0	5	.3846
7	118	6	0	1	0	0	0	1								2	.3333
8	144	7	1	0	1	1	1	0	0							4	.5714
9	236	12	0	0	0	1	0	1	0	0	0	0	0	0		2	.1667
10	184	9	1	0	0	1	0	0	1	0	1					4	.4444
11	137	7	0	0	1	0	0	0	0							1	.1429

Estimate the proportion of rabies caused by pet dogs. Also, place confidence limits on it.

Solution

We have in this case, N = 70 and n = 11. As in the foregoing examples, we prepare table 11.7.

Table 11.7 Certain intermediate calculations

Hospital	M_i	m_i	p_i	$\left(\frac{M_i - m_i}{M_i}\right)\frac{p_i q_i}{m_i - 1}$	M_i^2 Col (5)	$M_i^2(p_i - p_{m3})^2$
1	140	7	.4286	.0388	760.48	187.85
2	203	10	.3000	.0222	914.84	38.84
3	91	5	.4000	.0567	469.53	39.77
4	176	9	.3333	.0264	817.77	.21
5	121	6	.1667	.0264	386.52	393.78
6	263	13	.3846	.0187	1293.46	200.95
7	118	6	.3333	.0422	587.59	.09
8	144	7	.5714	.0388	804.56	1201.37
9	236	12	.1667	.0120	668.35	1498.00
10	184	9	.4444	.0294	995.37	437.68
11	137	7	.1429	.0194	364.12	661.96
Total	1813		3.6719	.3310	8062.59	4660.50

Since M_0 is not known, the estimate of proportion of rabies caused by pet dogs can, therefore, be obtained by using either of the two estimators in (11.33) and (11.40). In this example, we illustrate the use of estimator p_{m2} defined in (11.33). Thus,

$$p_{m2} = \frac{1}{n}\sum_{i=1}^{n} p_i$$

$$= \frac{3.6719}{11} \qquad \text{(from column (4) of table 11.7)}$$

$$= .3338$$

The estimate of variance of the above estimator is provided by (11.36) as

$$v(p_{m2}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{2b}^2 + \frac{1}{nN}\sum_{i=1}^{n}\left(\frac{M_i - m_i}{M_i}\right)\frac{p_i q_i}{m_i - 1}$$

where

$$s_{2b}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(p_i - p_{m2})^2$$

On using column (4) of table 11.7 and the value of p_{m2} calculated above, one obtains

$$s_{2b}^2 = \frac{1}{11-1}[(.4286-.3338)^2 + (.3000-.3338)^2 + ... + (.1429-.3338)^2]$$

$$= \frac{1}{10}[(.4286)^2 + (.3000)^2 + ... + (.1429)^2 - 11(.3338)^2]$$

$$= .0178$$

The value of second term in $v(p_{m2})$ is calculated by using column (5) of table 11.7. Thus we get

$$v(p_{m2}) = (\frac{1}{11} - \frac{1}{70})(.0178) + \frac{.3310}{(11)(70)}$$

$$= .001794$$

The confidence interval is given by the limits

$$p_{m2} \pm 2\sqrt{v(p_{m2})}$$

$$= .3338 \pm 2\sqrt{.001794}$$

$$= .3338 \pm .0847$$

$$= .2491, .4185$$

Thus had all the patients of rabies treated in 70 hospitals been contacted, the proportion of rabies caused by pet dogs would most probably have taken a value in the range of 24.91% to 41.85%. ■

Example 11.7

Using data of example 11.6, estimate the required proportion by using estimator p_{m3} given in (11.40). Also, work out the confidence interval for the parameter being estimated.

Solution

Again, we are given N = 70 and n = 11. The proportion of rabies caused by pet dogs is now estimated using estimator p_{m3}, where

$$p_{m3} = \frac{\sum_{i=1}^{n} M_i p_i}{\sum_{i=1}^{n} M_i}$$

Using columns (2) and (4) of table 11.7, the estimator p_{m3} is evaluated as

$$p_{m3} = \frac{1}{1813}[(140)(.4286) + (203)(.3000) + ... + (137)(.1429)]$$

$$= \frac{599.5844}{1813}$$

$$= .3307$$

The estimator of variance is provided by (11.43), where

$$v(p_{m3}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{3b}^2 + \frac{1}{nN\bar{M}^2} \sum_{i=1}^{n} \left(\frac{M_i^2(M_i - m_i)}{M_i}\right) \frac{p_i q_i}{m_i - 1}$$

As $\bar{M}$ is not available in this case, its sample estimate

$$\hat{\bar{M}} = \frac{1}{n} \sum_{i=1}^{n} M_i = \frac{1813}{11} = 164.82$$

will be used in its place. Thus, s_{3b}^2 can be put as

$$s_{3b}^2 = \frac{1}{\hat{\bar{M}}^2 (n-1)} \sum_{i=1}^{n} M_i^2 (p_i - p_{m3})^2$$

The use of column (7) of table 11.7 yields

$$s_{3b}^2 = \frac{4660.50}{(164.82)^2(10)} = .01716$$

The second component of $v(p_{m3})$ is evaluated by using total of column (6) of table 11.7. We thus obtain

$$v(p_{m3}) = \left(\frac{1}{11} - \frac{1}{70}\right)(.01716) + \frac{8062.59}{(11)(70)(164.82)^2}$$

$$= .0017$$

The required confidence limits, within which the population proportion of rabies caused by pet dogs is likely to fall, are arrived at by using

$$p_{m3} \pm 2\sqrt{v(p_{m3})}$$

$$= .3307 \pm 2\sqrt{.0017}$$

$$= .3307 \pm .0825$$

$$= .2482, .4132 \blacksquare$$

11.5 ESTIMATION OF MEAN / TOTAL USING PPSWR AND SRSWOR

It is possible to increase the efficiency of multistage sampling by making use of the auxiliary information that may be available for first and subsequent stage units. For instance, in a socio-economic survey where the ultimate sampling unit is the household, villages may be treated as psu's. These could be selected with probability proportional to size and with replacement, the size being the number of households, or population for the village. If the number of second stage units in the psu's differ considerably, it may be useful to select psu's with probability proportional to M_i values where M_i, as

before, is the number of ssu's in the i-th psu, i = 1, 2,...,N. One may also follow more efficient strategy for selecting ssu's by utilizing any other auxiliary information available for them in the selected psu's.

For discussion in this section, we assume that a sample of n psu's is selected WR using unequal probabilities $\{P_i\}$. From the M_i ssu's of the i-th selected psu, a sample of m_i ssu's is drawn with WOR simple random sampling. This sampling strategy gives following formulas.

Unbiased estimator of population total Y :

$$\hat{Y}_{mp} = \frac{1}{n} \sum_{i=1}^{n} M_i \frac{\bar{y}_i}{p_i} \tag{11.44}$$

Variance of the estimator $\hat{Y}_{mp}$:

$$V(\hat{Y}_{mp}) = \frac{1}{n} \sum_{i=1}^{N} \left(\frac{Y_i}{P_i} - Y \right)^2 P_i + \frac{1}{n} \sum_{i=1}^{N} \frac{M_i^2}{P_i} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) S_i^2 \tag{11.45}$$

where S_i^2 has been defined in (11.2).

Estimator of variance $V(\hat{Y}_{mp})$:

$$\begin{aligned} v(\hat{Y}_{mp}) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{M_i \bar{y}_i}{p_i} - \hat{Y}_{mp} \right)^2 \\ &= \frac{1}{n(n-1)} \left[\sum_{i=1}^{n} \left(\frac{M_i \bar{y}_i}{p_i} \right)^2 - n \hat{Y}_{mp}^2 \right] \end{aligned} \tag{11.46}$$

Corresponding estimator of population mean will be obtained by dividing $\hat{Y}_{mp}$ by M_o.

Estimator of population mean $\bar{Y}$:

$$\bar{y}_{mp} = \hat{Y}_{mp}/M_o = \frac{1}{nM_o} \sum_{i=1}^{n} \frac{M_i \bar{y}_i}{p_i} \tag{11.47}$$

Variance of the estimator $\bar{y}_{mp}$:

$$V(\bar{y}_{mp}) = \frac{V(\hat{Y}_{mp})}{M_o^2} \tag{11.48}$$

Estimator of variance $V(\bar{y}_{mp})$:

$$v(\bar{y}_{mp}) = \frac{v(\hat{Y}_{mp})}{M_o^2} \tag{11.49}$$

where $V(\hat{Y}_{mp})$ and $v(\hat{Y}_{mp})$ are given in (11.45) and (11.46) respectively.

Example 11.8

In order to estimate the total production of wheat in a certain district, 16 villages out of a total of 410 villages were selected using PPS with replacement sampling, the size measure being the net cropped area in hectares. The total net cropped area for the district was 140576 hectares. About 3 percent of farmers were selected from the sampled villages. The total produce of wheat (in quintals) for each of the sampled farmer, as reported by him, was recorded. These data are presented in table 11.8.

Table 11.8 Net cropped area (x_i) for the village and production of wheat (y_{ij}) for the selected farmers

Village	x_i	$p_i = \frac{x_i}{X}$	M_i	m_i	y_{ij}				Total ($y_{i.}$)	$\bar{y}_i = \frac{y_{i.}}{m_i}$
1	420	.00299	96	3	138	166	190		494	164.67
2	613	.00436	112	3	142	185	215		542	180.67
3	178	.00127	40	1	110				110	110.00
4	199	.00142	46	1	133				133	133.00
5	345	.00245	86	3	160	164	210		534	178.00
6	467	.00332	122	4	100	162	85	124	471	117.75
7	123	.00087	42	1	107				107	107.00
8	328	.00233	92	3	140	163	116		419	139.67
9	150	.00107	61	2	105	98			203	101.50
10	764	.00543	190	6	200	140	173	160	902	150.33
					101	128				
11	269	.00191	76	2	120	135			255	127.50
12	483	.00344	138	4	149	113	161	131	554	138.50
13	389	.00277	98	3	110	124	90		324	108.00
14	212	.00151	66	2	190	105			295	147.50
15	160	.00114	34	1	166				166	166.00
16	532	.00378	150	5	136	170	100	156	702	140.40
					140					

Estimate the total production of wheat in the district, and determine confidence limits for it.

Solution

We are given that X = 140576, N = 410, and n = 16. In table 11.8, the observations in column (7) are the total of observations in column (6) for a given village. The estimate of total production of wheat is given by (11.44) as

$$\hat{Y}_{mp} = \frac{1}{n} \sum_{i=1}^{n} \frac{M_i \bar{y}_i}{p_i}$$

$$= \frac{1}{16}\left[\frac{(96)\,(164.67)}{.00299} + \frac{(112)\,(180.67)}{.00436} + \ldots + \frac{(150)\,(140.40)}{.00378}\right]$$

$$= \frac{81422932}{16}$$

$$= 5088933.2$$

We now obtain estimate of variance $V(\hat{Y}_{mp})$. For this we use (11.46). Therefore,

$$v(\hat{Y}_{mp}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{M_i \bar{y}_i}{p_i} - \hat{Y}_{mp}\right)^2$$

$$= \frac{1}{(16)\,(15)}\left[\left\{\frac{(96)\,(164.67)}{.00299} - 5088933.2\right\}^2 + \left\{\frac{(112)\,(180.67)}{.00436}\right.\right.$$

$$\left.\left. - 5088933.2\right\}^2 + \ldots + \left\{\frac{(150)\,(140.40)}{.00378} - 5088933.2\right\}^2\right]$$

$$= \frac{1.00376 \times 10^{13}}{(16)\,(15)}$$

$$= 4.18236 \times 10^{10}$$

The confidence limits, within which the total production of wheat in the district is likely to fall, are

$$\hat{Y}_{mp} \pm 2\sqrt{v(\hat{Y}_{mp})}$$

$$= 5088933.2 \pm 2\sqrt{4.18236 \times 10^{10}}$$

$$= 5088933.2 \pm 409016.4$$

$$= 4679916.8,\ 5497949.6$$

Thus the desired confidence interval is [4679916.8, 5497949.6] quintals. ■

The reader will notice that in the above example only one second stage unit (farmer) was selected from some of the selected first stage units (villages). It is possible to estimate the variance of the estimator of population total, or mean, in such a case where the sample of psu's is selected with replacement. In case the psu's are selected without replacement, one needs a sample of at least two second stage units from each selected psu to get an estimator of variance.

11.6 SOME FURTHER REMARKS

11.1 The results of this chapter can be easily extended to more than two stages of sampling, or stratified sampling. For these, and various other variants, reader may refer to Sukhatme *et al.* (1984) for details.

11.2 In case the primary stage units are selected with PPS with replacement and if $\hat{y}_{i.}$ denotes an unbiased estimator of the i-th sample psu total $y_{i.}$, i = 1, 2, ..., n, then an unbiased estimator of population total is given by

$$\hat{Y}_{mp} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{y}_{i.}}{p_i} \tag{11.50}$$

Also, an unbiased estimator of variance $V(\hat{Y}_{mp})$ is given by

$$v(\hat{Y}_{mp}) = \frac{1}{n(n-1)} \left[\sum_{i=1}^{n} \left(\frac{\hat{y}_{i.}}{p_i} \right)^2 - n\hat{Y}_{mp}^2 \right] \tag{11.51}$$

The results in (11.50) and (11.51) hold for any number of stages and all kinds of estimators used at different stages, in building the estimator $\hat{y}_{i.}$.

LET US DO

11.1 Define multistage sampling, and identify the situations where it is to be preferred over usual SRS for selecting a sample of ultimate units.

11.2 Suppose the objective of a survey is to estimate the total wheat production in a state. The state has a number of districts and each district has several development blocks. Many villages constitute a block while there are several farmers in each village. The frame of villages is also available at the state level. Suggest what should be the psu's and ssu's, if one is to go for a two-stage sampling design ?

11.3 Give expression for the unbiased estimator of mean for a two-stage sampling design. What will be its variance, and how will you estimate it ? Choose your own sampling schemes at both the stages of sampling.

11.4 A district is running 110 nursery schools (*anganwadi*) in the rural area. The total number of children in these schools, is known to be 8040. The Department of Foods and Nutrition of a university has undertaken a project to determine the quality of food intake by the children in these schools. For this purpose, a sample of 10 schools was selected using WOR equal probability sampling. From each selected school, about 5 percent of children were selected using same procedure. Elaborate diet records were kept for each selected child, and the average daily calory intake was then determined. The information thus collected, is presented in the following table.

School	M_i	m_i	Average calory intake for sample children				
1	100	5	650	680	712	766	770
2	93	5	745	690	703	740	671
3	40	2	736	692			
4	80	4	703	777	687	714	

Table continued ...

School	M_i	m_i	Average calory intake for sample children				
5	42	2	699	724			
6	62	3	760	704	680		
7	96	5	715	660	690	670	650
8	46	2	668	701			
9	99	5	704	712	692	697	687
10	104	5	714	716	704	725	730

Estimate unbiasedly the average daily calory intake for the children of all the 110 schools, and also obtain the confidence interval for it.

11.5 In case the total number of second stage units in the population is not known, discuss how will you estimate the population mean ? Also, give the expressions for the variance and estimator of variance for the estimator(s) you propose to use.

11.6 In addition to the crop cutting survey, the Department of Agriculture had decided to estimate per hectare wheat yield by taking cultivator as the ultimate sampling unit. The list of all the 300 villages comprising the district was available. However, information regarding the number of cultivators in each village was not available. Two-stage sampling was, therefore, thought to be an appropriate design. A WOR simple random sample of 10 villages was drawn. About 10 percent of the cultivators in each sample village were selected using same sampling scheme. The per hectare wheat yield (in quintals) obtained by the selected cultivators was asked for by the investigator. This information, along with the total number of cultivators (M_i) and the number of cultivators sampled (m_i), is presented in the following table :

Village	M_i	m_i	Per hectare wheat yield				
1	40	4	29.70	32.80	28.80	30.65	
2	28	3	33.18	32.60	33.41		
3	35	4	27.76	25.30	31.40	30.60	
4	24	2	34.17	31.69			
5	17	2	27.28	32.40			
6	30	3	35.20	33.40	29.50		
7	25	3	28.24	27.45	34.90		
8	21	2	30.45	29.12			
9	38	4	28.23	31.56	33.45	29.70	
10	50	5	33.60	28.70	32.07	32.65	28.40

Estimate average per hectare yield of wheat in the district using estimator $\bar{y}_{m2}$ in (11.10). Also, build up the confidence interval for this average.

11.7 The excessive rains have caused damage to cotton crop in a certain area consisting of 132 villages. Before deciding on the extent of relief to be given to cultivators, the administration has decided to estimate the average per hectare loss for the area. For this purpose, a sample of 10 villages was selected using without replacement SRS, and about 5 percent of cultivators from the sample villages were selected using the same sampling scheme. Per hectare loss (in '00 rupees) incurred by each selected cultivator was assessed by visiting his fields. The results are as follows :

Village	M_i	m_i	Loss per hectare				
1	43	2	8	12			
2	76	4	16	7	11	13	
3	46	2	14	9			
4	67	3	17	10	12		
5	97	5	14	10	15	17	11
6	75	4	16	7	9	14	
7	83	4	9	13	11	13	
8	54	3	15	15	13		
9	69	3	12	9	11		
10	58	3	14	15	14		

Work out the estimated average per hectare loss, and also construct the confidence interval for it.

11.8 A project has been undertaken to study the feeding and rearing practices of sheep in Rajasthan state of India. As a first step, a survey was conducted to estimate the total sheep population in the state. For this purpose, a WOR simple random sample of 12 development blocks, from a total of 200 blocks, was selected. About 3 percent of the villages in the sample blocks were chosen using same sampling scheme. The following table presents the number of sheep in the selected villages along with the total number of villages (M_i) and the number of selected villages (m_i), in the sample blocks.

Block	M_i	m_i	Number of sheep				
1	160	5	141	376	267	201	55
2	130	4	60	228	270	80	
3	128	4	177	95	265	301	
4	114	3	225	107	119		
5	70	2	80	101			
6	105	3	201	69	134		
7	89	3	135	240	405		
8	71	2	309	111			
9	97	3	250	280	170		
10	106	3	96	269	118		
11	128	4	314	246	107	80	
12	69	2	307	286			

Using estimator $\hat{Y}_{m2}$ in (11.23), estimate the total number of sheep in the state. Also work out the standard error of the estimate, and construct the confidence interval for this total. The total number of villages in the state is 24000.

11.9 From the sample data given in exercise 11.8, estimate total number of sheep in the state unbiasedly. Also, work out the standard error of your estimate.

11.10 A soap factory is planning to use *neem (Azadirachta indica)* oil in manufacturing a new brand of toilet soap. For extracting oil, the *neem* seeds are to be collected at 100 centers established for the purpose. It is felt that the persons from villages, falling within a radius of 20 km of a *neem* seed collection center, will find it remunerative to sell *neem* seeds to the collection centers. For getting an idea about the *neem* seed supply position, the factory wants to estimate the total number of *neem* trees in the procurement area for all the 100 collection centers. For this purpose, an SRS without replacement sample of 10 centers was selected. Using same sampling procedure, about 10 percent of villages falling within the 20 km radius of the sample centers were selected as second stage units. Number of *neem* trees in the selected villages are given below along with the values of M_i and m_i.

Center	M_i	m_i	No. of *neem* trees					
1	30	3	40	61	76			
2	44	4	88	25	49	33		
3	56	6	75	32	56	44	37	55
4	36	4	42	25	34	39		
5	27	3	33	47	29			
6	51	5	55	43	61	40	36	
7	42	4	27	42	36	58		
8	30	3	60	46	39			
9	28	3	51	37	42			
10	33	3	64	34	56			

Estimate the total number of *neem* trees falling in the procurement area for all the 100 centers by using an appropriate estimator. Also, place confidence limits on this total.

11.11 Give an unbiased estimator for population proportion for a two-stage sampling design. Also, write expressions for its variance and estimator of variance.

11.12 The objective of a study was to estimate the proportion of liquor shops in a state selling spurious liquor. The state consists of 117 development blocks and has 8068 liquor shops. A WOR simple random sample of 10 blocks was selected for this study. About 10 percent of liquor shops were selected from the sampled development blocks. The liquor in the selected shops was examined. The collected information, along with the total number of liquor shops in the sample blocks (M_i) and the number of shops selected (m_i), is given as follows.

Block	:	1	2	3	4	5	6	7	8	9	10
M_i	:	104	96	60	102	101	91	87	90	76	89
m_i	:	10	10	6	10	10	9	9	9	8	9
Shops selling spurious liquor	:	4	5	4	6	7	5	4	3	5	6

Estimate unbiasedly the proportion of shops selling spurious liquor in the state. Also, work out the standard error of your estimate, and place confidence limits on the value of the parameter under study.

11.13 There are 76 senior citizen homes (SCH's) in a state which care for the old persons who are unable to stay at their native homes due to some reasons. It is felt that once a person is admitted to an SCH and stays there for sometime, he/she does not like to return to his/her native home even if the circumstances that had compelled him/her to live in the SCH have changed and are favorable for his/her return. For verifying this belief, a WOR simple random sample of 9 SCH's was drawn. About 10 percent of the persons living in the sample homes were interviewed. The information, so collected, is given in the following table, where 1 indicates that the person is not willing to return to his/her native home, and 0 otherwise. The total number of old persons staying in a sample SCH (M_i), and the number of persons selected for interview (m_i), are also given in the table.

SCH	M_i	m_i	Scores
1	110	11	1 1 0 1 1 1 0 0 1 1 1
2	36	4	1 1 1 0
3	70	7	0 0 1 1 1 1 1
4	82	8	1 1 1 0 1 1 0 1
5	66	7	1 1 0 0 0 1 1
6	97	10	1 1 1 0 1 1 1 1 1 0
7	85	9	0 1 1 1 1 1 0 1 1
8	67	7	1 1 1 1 1 0 1
9	56	6	1 0 0 0 1 1

Using estimator p_{m2} in (11.33), estimate the proportion of persons not willing to return to their native homes. Also, work out the standard error of the estimate, and place confidence limits on the population value.

11.14 A district transport office is concerned about the proportion of tractors plying without valid documents in the district. The total number of tractors operating in the district is not exactly known. This is because some of the tractors now plying in the district might have been registered in some other districts/states, whereas some other tractors registered with the office might have been sold outside the district. It was, therefore, thought appropriate to use two-stage sampling design to estimate this proportion. For this purpose, 16 villages out of the total of 400 villages in the district, were selected using SRS without replacement. About 20 percent of tractors in the sample villages were selected

using same sampling scheme. Documents of the selected tractors were then examined. The information thus collected is presented below, along with the total number of tractors in sample villages (M_i) and number of tractors examined (m_i) for validity of documents. The abbreviation TWD in the following table stands for tractors without valid documents.

Village	M_i	m_i	TWD	Village	M_i	m_i	TWD
1	40	8	6	9	36	7	5
2	13	3	3	10	27	5	3
3	36	7	5	11	15	3	2
4	48	10	7	12	62	12	9
5	32	6	4	13	33	7	5
6	42	8	7	14	44	9	6
7	60	12	8	15	38	8	6
8	28	6	4	16	55	11	7

Using estimator p_{m3} in (11.40), estimate the proportion of tractors in the district plying without valid documents. Also, place confidence limits on this proportion.

11.15 A with replacement PPS sample of 10 wards out of a total of 63 wards comprising a town was selected, the size measure being the number of households in a ward available from the population census records. From the sample wards, about 5 percent of households were selected through WOR simple random sampling. The per month expenses (in rupees) on vegetables were recorded for the selected households. This information is given in the table below along with the values of M_i and m_i which carry their usual meaning. The total number of households in all the 63 wards is 12800.

Ward	M_i	m_i	Total monthly expenses for m_i households
1	160	8	1600
2	240	12	2000
3	218	11	2350
4	148	7	1760
5	276	14	3500
6	238	12	3320
7	196	10	3035
8	256	13	2600
9	217	11	3730
10	177	9	3100

What is the estimated per month expenditure on vegetables for a household ? Work out the standard error of your estimate, and place confidence limits on the population value.

CHAPTER 12

Sampling from Mobile Populations

12.1 INTRODUCTION

The estimation of size is of immense importance in a variety of mobile biological populations. It helps to study population growth, ecological adaptation, natural selection, evolution, maintenance of many *wildlife populations*, and so on. Unlike other populations considered in previous chapters, the sampling units in the wildlife populations do not remain fixed at one place. They are rather highly mobile. Therefore, for mobile populations, it is essential to use alternative approach for sampling and estimation.

One of the commonly used procedures is the *capture-recapture method.* In this method, animals, birds, reptiles, etc., are captured by any appropriate device, viz., by using drugs, snares, dart-gun, pitfalls, box traps, net traps, enclosure traps, nest traps, etc., and are marked in a variety of ways. Bands or rings may be used for birds and bats, ear tags and collars for mammals, whereas fish and reptiles could be marked using jawtags.

This basic technique seems to have been used first by Petersen (1896) to estimate a plaice population, and then by Lincoln (1930) to estimate the total number of ducks in North America. The same method was adopted independently by Jackson (1933), who used it for estimating the true density of tsetse flies. The procedure has been further discussed in detail by Bailey (1951) and Chapman (1951 and 1952).

As mentioned above, the mobile population is a wider term which includes animals, birds, reptiles, insects, etc. In what follows, we shall sometimes use the word 'animals' to refer to all types of wildlife populations.

The estimation of population using capture-recapture principle rests on certain *assumptions*. These are listed below :

1. The population is closed. It means the mortality, recruitment, and migration during the period of data collection are negligible.
2. All animals have the same probability of being caught.
3. Marking does not affect the catchability of an animal.
4. Animals tend to be distributed randomly and redistribute themselves at random after release.
5. Animals do not loose their marks in the time between the two samples.

Basically, the procedure involves following steps :

1. A random sample of t units is drawn from a closed wildlife population of interest of size, say, N.
2. The sampled t animals are marked for future identification and then released into the population.

3. After allowing a time for marked and unmarked animals to mingle, another random sample of size, say, n units is drawn from the same population, and the marked animals are counted. The proportion of marked animals in the population is obviously t /N = P. This implies

$$N = \frac{t}{P} \tag{12.1}$$

Here t is known. The proportion P can be estimated by p, the proportion of marked animals in the second sample.

If the random sample in step (3) is of fixed size, the whole of the procedure is termed as *direct sampling*. In case, the size of the sample in step (3) is not fixed, and the investigator goes on sampling until a fixed number of marked animals is observed, the procedure is known as *inverse sampling*. First we discuss the direct sampling approach.

12.2 ESTIMATION OF POPULATION SIZE USING DIRECT SAMPLING

In the direct sampling procedure, the steps (1) and (2) are same as in section 12.1. In step (3), the second sample of n animals is drawn and marked animals are counted. If s be the number of marked animals observed in the second sample, the proportion of marked animals in sample of size n is given by

$$p = \frac{s}{n} \tag{12.2}$$

The model is known as *hypergeometric* or *binomial* according as the second random sample of fixed size in step (3) is drawn WOR or with replacement respectively.

12.2.1 *Hypergeometric Model*

Chapman (1951) has shown that the estimator $\hat{N}_p$, defined in (12.6) and proposed earlier by Petersen (1896) for the binomial model, is biased and the bias can be large for small samples. He considers a less biased and more efficient estimator $\hat{N}_c$ for the population size N, when the second sample of n animals in step (3) is drawn without replacement.

Chapman's estimator of population size N :

$$\hat{N}_c = \frac{(t+1)(n+1)}{s+1} - 1 \tag{12.3}$$

Approximate variance of estimator $\hat{N}_c$:

$$V(\hat{N}_c) = N^2 (\mu^{-1} + 2\mu^{-2} + 6\mu^{-3}) \tag{12.4}$$

where $\mu = tn/N$.

Estimator of variance $V(\hat{N}_c)$:

$$v(\hat{N}_c) = \frac{(t+1)(n+1)(t-s)(n-s)}{(s+1)^2 (s+2)} \tag{12.5}$$

If $t+n \geq N$, the estimator $\hat{N}_c$ and $v(\hat{N}_c)$ are unbiased.

Example 12.1

A large number of field rats are inhabiting a state sugarcane farm and are causing severe damage to the sugarcane crop. The farm officials are concerned about it. Before applying the suitable insecticide, the officials wish to estimate the total population of rats on the farm. They trapped 124 rats at random locations. The rats were marked and then released at scattered points at the farm. After three days 256 rats were trapped from the farm out of which 24 were found to be marked. Treating the 256 rats as a WOR random sample, estimate the rat population using Chapman's estimator $\hat{N}_c$ in (12.3), and determine confidence limits for N.

Solution

We are given that t=124, n=256, and s=24. The expression for the required estimator, from (12.3), is

$$\hat{N}_c = \frac{(t+1)(n+1)}{s+1} - 1$$

On substituting different values, we get the estimate of rat population as

$$\hat{N}_c = \frac{(124+1)(256+1)}{(24+1)} - 1$$
$$= 1284$$

The estimator of variance is given by (12.5). Using it, we obtain $v(\hat{N}_c)$ as

$$v(\hat{N}_c) = \frac{(t+1)(n+1)(t-s)(n-s)}{(s+1)^2(s+2)}$$
$$= \frac{(124+1)(256+1)(124-24)(256-24)}{(24+1)^2(24+2)}$$
$$= 45864.61$$

The required confidence limits are computed by using

$$\hat{N}_c \pm 2\sqrt{v(\hat{N}_c)}$$
$$= 1284 \pm 2\sqrt{45864.61}$$
$$= 855.68,\ 1712.32$$
$$\approx 856,\ 1712$$

The officials could be reasonably confident that the true rat population in the sugarcane farm under study will be in the interval [856, 1712]. ■

12.2.2 *Binomial Model*

When the second sample of n units drawn in step (3) is a WR sample, the theoretical framework for the procedure could be developed using binomial model. For this situation, we consider two estimators - one due to Petersen (1896), and the other due to Bailey (1951). Substituting the value of p from (12.2) in place of P in (12.1), one gets the maximum likelihood estimator of N, also called *Petersen estimator.*

Petersen estimator of population size N :

$$\hat{N}_p = \frac{nt}{s} \tag{12.6}$$

Estimator of variance $V(\hat{N}_p)$:

$$v(\hat{N}_p) = \frac{nt^2(n-s)}{s^3} \tag{12.7}$$

Example 12.2

Adams (1957) conducted a survey to estimate Snowshoe Hares population on an island in Flathead Lake, Montana, where immigration and emigration could not occur. The birth rate during winter was also negligible. Mortality was assumed to be same in the marked and unmarked individuals. During January, 27 hares were captured, marked, and released into the population. In the following March, 23 hares were recaptured of which 17 were found marked. Using (12.6), estimate the total number of Snowshoe Hares, and set confidence limits for it.

Solution

We are given t = 27, n = 23, and s = 17. From (12.6),

$$\hat{N}_p = \frac{nt}{s} = \frac{(23)\ (27)}{17} = 36.53 \approx 37$$

is the estimated number of Snowshoe Hares in the area under study. The estimate of variance of $\hat{N}_p$ is obtained by using (12.7). Thus,

$$v(\hat{N}_p) = \frac{nt^2(n-s)}{s^3}$$

$$= \frac{(23)\ (27)^2\ (23-17)}{(17)^3}$$

$$= 20.48$$

Then using estimated variance $v(\hat{N}_p)$ obtained above, we work out confidence limits through (2.8). These limits are

$$\hat{N}_p \pm 2\sqrt{v(\hat{N}_p)}$$

$$= 36.53 \pm 2\sqrt{20.48}$$

$$= 27.48, 45.58$$

$$\approx 27, 46$$

The confidence limits obtained above indicate that the actual total number of Snowshoe Hares, in the area surveyed, would probably be between 27 and 46. ■

The estimator $\hat{N}_p$ is biased and the expression for its variance is difficult to obtain. For estimator $\hat{N}_p$ in (12.6), we assume that n is sufficiently large, so that, s takes value greater than zero. Bailey (1951) suggested a slightly adjusted estimator which is relatively less biased. The expression for the variance of this estimator is also difficult to find. However, nearly unbiased estimator of variance for sample sizes not too small, is available.

Bailey's estimator of population size N :

$$\hat{N}_b = \frac{t(n+1)}{s+1} \tag{12.8}$$

Approximately unbiased estimator of variance $V(\hat{N}_b)$:

$$v(\hat{N}_b) = \frac{t^2(n+1)(n-s)}{(s+1)^2(s+2)} \tag{12.9}$$

Example 12.3

In the Harike (Punjab) wetland covering an area of about 40 sq km, the local duck population is 1500. During October to February, ducks from Siberia, China, and Europe migrate to Harike bird sanctuary. An ornithologist wishes to estimate the total duck population during this period so that he is able to estimate the number of migratory ducks. The survey procedure was initiated at the end of December. A random sample of 210 ducks was taken. The ducks captured were tagged and released into the population. Sufficient time was allowed so that the tagged ducks get mixed with the untagged ones. Then another WR random sample of 700 ducks was observed using binoculars. Twenty of them were found tagged. Estimate the total duck population using Bailey's estimator $\hat{N}_b$ in (12.8), and place confidence limits on N.

Solution

From the statement of the example, we have t = 210, s = 20, and n = 700. The required estimate is computed using (12.8). Thus,

$$\hat{N}_b = \frac{t(n+1)}{s+1} = \frac{210(700+1)}{20+1} = 7010$$

The estimate of variance of $\hat{N}_b$ is given by (12.9). We, therefore, write

$$v(\hat{N}_b) = \frac{t^2(n+1)(n-s)}{(s+1)^2(s+2)}$$

On making substitutions, one gets

$$v(\hat{N}_b) = \frac{(210)^2(700+1)(700-20)}{(20+1)^2(20+2)}$$

$$= 2166727.2$$

We now require confidence limits. These are given by

$$\hat{N}_b \pm 2\sqrt{v(\hat{N}_b)}$$

$$= 7010 \pm 2\sqrt{2166727.2}$$

$$= 4066.04, 9953.96$$

$$\approx 4066, 9954$$

The survey concludes that the estimate of duck population in Harike wetland is 7010. However, the actual duck population is expected to be in the range of 4066 to 9954 with probability approximately .95. ■

Example 12.4

Odum and Pontin (1961) undertook a survey to estimate the population of yellow ant (*Lasius flavus*) at place "Bowling alley" located in Wytham Woods near Oxford, England. These ants normally do not venture above ground. The population was constant in size throughout the experiment as (1) the individuals were relatively long lived, (2) the newly emerged individuals from the pupae could be recognized by their lighter color and, therefore, eliminated in the tagging and recapture counts, and (3) there was negligible or no movement of individuals from one colony to another. A flat stone was placed on the top of the mound. The ants constructed extensive galleries under the stone, where warmth from the sun was favorable for the development of the pupae. The stone was lifted gently and 500 individuals were removed, marked with radioactive P^{32} solution and allowed to crawl back under the stone on their own accord. Mortality during the marking process was negligible. After allowing a period of two days for a proper mixing of marked and unmarked ones, the stone was again lifted and a random WR recapture sample of 437 ants was taken. In this sample 68 ants were found to be marked. Estimate the ant population in the colony using (12.8), and place confidence limits on it.

Solution

In the statement t = 500, n = 437, and s = 68. From (12.8), we have

$$\hat{N}_b = \frac{t(n+1)}{s+1} = \frac{500(437+1)}{68+1} = 3173.9 \approx 3174$$

as the estimated number of yellow ants at Bowling alley. The variance estimator for $\hat{N}_b$ is provided by (12.9). Thus,

$$v(\hat{N}_b) = \frac{t^2(n+1)(n-s)}{(s+1)^2(s+2)}$$

On making substitutions, one gets

$$v(\hat{N}_b) = \frac{(500)^2(437+1)(437-68)}{(68+1)^2(68+2)}$$

$$= 121239.5$$

We now work out confidence limits following (2.8). These are

$$\hat{N}_b \pm 2\sqrt{v(\hat{N}_b)}$$

$$= 3173.9 \pm 2\sqrt{121239.5}$$

$$= 2477.5, 3870.3$$

$$\approx 2478, 3870$$

This indicates that the survey has provided a strong support to the idea that the true underground ant population in the colony is likely to be in the range of 2478 to 3870. ■

12.3 ESTIMATION OF POPULATION SIZE USING INVERSE SAMPLING

Inverse sampling is an alternative procedure to estimate the size of the population. The model differs from direct sampling in the sense that s, the number of tagged animals to be observed, is fixed in advance. The size of the sample in step (3) is, therefore, a random variable, since we continue to select animals till s tagged animals are observed. Again, the random sampling in step (3) may be without or with replacement. For WOR sampling, the model will be called *negative hypergeometric* and in case of WR it is termed as *negative binomial* model.

12.3.1 *Negative Hypergeometric Model*

For WOR sampling, we shall consider the modified maximum likelihood estimator due to Bailey (1951). This estimator is unbiased for the population size, and exact expression for its variance is available. The properties of this estimator have been discussed in detail by Chapman (1952).

Bailey's unbiased estimator of population size N:

$$\hat{N}_{ib} = \frac{n(t+1)}{s} - 1 \tag{12.10}$$

Variance of the estimator $\hat{N}_{ib}$:

$$V(\hat{N}_{ib}) = \frac{(N+1)(N-t)(t-s+1)}{s(t+2)} \tag{12.11}$$

Unbiased estimator of variance $V(\hat{N}_{ib})$:

$$v(\hat{N}_{ib}) = \frac{t-s+1}{(t+1)(s+1)} [\hat{N}_{ib}^2 - \hat{N}_{ib}(t-1) - t] \tag{12.12}$$

On the average, the inverse sampling method is little more efficient than the direct method for a given coefficient of variation. The inverse method provides an estimator of N with an expected sample size

$$E(n \mid t, s) = \frac{(N+1)s}{t+1}$$

which is smaller than the sample size required for the direct method, though the difference is usually small. However, if the investigator has absolutely no knowledge about N, the expected sample size may be very large with an improper choice of t and s. This increases unpredictability of what n will actually turn out to be.

Example 12.5

A fish farmer wishes to estimate the total number of fish in his tank. To arrive at the estimate, he selects 240 fish at random. These are returned to the tank with jaw-tags. Some time is allowed for them to distribute uniformly in the tank. At a later date another WOR random sample is drawn until 22 tagged fish are recaptured. In all, 440 fish had to be captured to find 22 tagged fish. Estimate N, and also work out confidence interval for it.

Solution

From the above statement, we have t = 240, s = 22, and n = 440. This is the case where negative hypergeometric distribution model holds. The estimate of total count of fish in the tank is obtained by using Bailey's estimator $\hat{N}_{ib}$. Thus from (12.10),

$$\hat{N}_{ib} = \frac{n(t+1)}{s} - 1$$

$$= \frac{440(240+1)}{22} - 1$$

$$= 4819$$

The estimate of variance of $\hat{N}_{ib}$ is provided by (12.12). The expression is

$$v(\hat{N}_{ib}) = \frac{t-s+1}{(t+1)(s+1)} [\hat{N}_{ib}^2 - \hat{N}_{ib}(t-1) - t]$$

On making substitutions, we get

$$v(\hat{N}_{ib}) = \frac{240-22+1}{(240+1)(22+1)} [(4819)^2 - (4819)(240-1) - 240]$$

$$= 872000.9$$

The confidence limits are computed following (2.8) as

$$\hat{N}_{ib} \pm 2\sqrt{v(\hat{N}_{ib})}$$

$$= 4819 \pm 2\sqrt{872000.9}$$

$$= 2951.4, 6686.6$$

$$\approx 2951, 6687$$

The farmer estimates the total number of fish as 4819. Further, he is reasonably confident that the true fish population in the tank could be expected in the range of 2951 to 6687. ■

12.3.2 *Negative Binomial Model*

We consider once again Petersen estimator $\hat{N}_p = tn/s$. The estimator has been examined in detail by Chapman (1952). For this model, the estimator $\hat{N}_p$, denoted by $\hat{N}_{ip}$ in (12.13), is unbiased for N and has an exact variance expression. The necessary formulas corresponding to this estimator are given in the following box.

Unbiased estimator of population size N :

$$\hat{N}_{ip} = \frac{nt}{s} \tag{12.13}$$

Variance of estimator $\hat{N}_{ip}$:

$$V(\hat{N}_{ip}) = \frac{N(N-t)}{s} \tag{12.14}$$

Estimator of variance $V(\hat{N}_{ip})$:

$$v(\hat{N}_{ip}) = \frac{nt^2(n-s)}{s^2(s+1)} \tag{12.15}$$

The expressions (12.6) and (12.7) hold only for $s > 0$. However, this constraint offers no difficulty in case of negative binomial and negative hypergeometric models since s is fixed in advance. As in sampling WOR, the inverse sampling method is slightly more efficient than the direct method but it also suffers from the drawback that expected sample size may be very large if improper choice of t and s is made.

Example 12.6

Wildlife biologists are interested in estimating the number of squirrels inhabiting a zoo area. A random sample of 96 squirrels was trapped, ear-tagged, and released into the population at scattered places so that they get mixed with the unmarked squirrels uniformly. After 2 days, the zoo area was combed with binoculars till 24 marked squirrels were found . In all, the investigator had to observe 252 squirrels to locate 24 marked ones. Estimate the true population of squirrels, and construct confidence interval for the total number N.

Solution

We have $t = 96$, $n = 252$, and $s = 24$. The problem is related to negative binomial model. We, therefore, use Petersen estimator $\hat{N}_{ip}$ given in (12.13). Thus,

$$\begin{aligned}\hat{N}_{ip} &= \frac{nt}{s} \\ &= \frac{(252)\,(96)}{24} \\ &= 1008\end{aligned}$$

is the estimated number of squirrels inhabiting the zoo area. The estimate of variance of $\hat{N}_{ip}$ is provided by (12.15). This is

$$\begin{aligned}v(\hat{N}_{ip}) &= \frac{nt^2\,(n-s)}{s^2\,(s+1)} \\ &= \frac{(252)\,(96)^2\,(252-24)}{(24)^2\,(24+1)} \\ &= 36771.8\end{aligned}$$

We now work out the confidence limits. These are given by

$$\begin{aligned}&\hat{N}_{ip} \pm 2\sqrt{v(\hat{N}_{ip})} \\ &= 1008 \pm 2\sqrt{36771.8} \\ &= 624.5,\ 1391.5 \\ &\approx 624,\ 1392\end{aligned}$$

The limits of the confidence interval computed above indicate that the actual total number of squirrels inhabiting the zoo area is likely to be in the range of 624 to 1392. ■

12.4 DETERMINING THE SAMPLE SIZES

An appropriate definition of the accuracy of an estimator is due to Robson and Regier (1964). Let (1-α) be the probability that the Petersen estimator $\hat{N}_p$ will not differ from the true population size N by more than 100A percent. Then

$$1-\alpha \leq \text{Prob}\left(-A < \frac{\hat{N}_p - N}{N} < A\right)$$

where α and A are to be chosen by the experimenter. Keeping α =.05, three levels of A are proposed :

1. A = .50, for obtaining rough idea of population size.
2. A = .25, for reasonably accurate estimation of N.
3. A = .10, for critical observation of population dynamics.

Assuming N > 100, Robson and Regier (1964) use normal approximation to the hypergeometric distribution. Given α and A, they solve

$$1-\alpha = \phi\left(\frac{A\sqrt{D}}{1-A}\right) - \phi\left(\frac{-A\sqrt{D}}{1+A}\right) \tag{12.16}$$

for D, where φ(z) is the cumulative standard normal distribution. From Robson and Regier (1964, table 2), values of D satisfying equation (12.16) for selected α and A are given in table 12.1.

Table 12.1 Values of D for given α and A

1-α	A	D
.75	.50	4.75
.90	.50	14.8
.90	.25	45.5
.95	.50	24.4
.95	.25	69.9
.95	.10	392
.99	.10	695

It has also been shown that by using an estimate of N, one can obtain different possible pairs of sample sizes t and n satisfying the constraint

$$\frac{tn\,(N-1)}{(N-t)\,(N-n)} = D \tag{12.17}$$

All the pairs of sample sizes t and n satisfying (12.17) will, however, not be optimum. We now consider the problem of determining optimum sample sizes for fixed total budget, and when the total cost is minimized for the given accuracy of the estimator.

12.4.1 *Optimum Sample Sizes for Fixed Cost*

Let c_1 be the cost of catching and marking an animal, and c_2 the cost per animal of catching and examining it. If c_0 is the overhead cost, a reasonable *cost function* considered by Robson and Regier (1964) is given by

$$C = c_0 + tc_1 + nc_2 \tag{12.18}$$

They also show that, for the given budget at disposal and predetermined accuracy A, an optimum choice of t and n which minimizes α is given by

$$\frac{c_1 t}{c_2 n} = \frac{N-n}{N-t} \tag{12.19}$$

For given values of c_1 and c_2, the equations (12.18) and (12.19) can be solved for t and n, using a rough guess value of N.

Optimal t and n for given budget C, accuracy A, and rough guess of N :

$$t = (C - c_0 - c_2 n)/c_1 \tag{12.20}$$

$$n = K_1 \pm K_1 K_2 \tag{12.21}$$

where

$$\left.\begin{aligned} K_1 &= \frac{Nc_1 - C + c_0}{c_1 - c_2} \\ K_2 &= \left[\frac{c_1 (Nc_2 - C + c_0)}{c_2 (Nc_1 - C + c_0)}\right]^{\frac{1}{2}} \end{aligned}\right] \tag{12.22}$$

An admissible value of n is found from (12.21), and substituted in (12.20) to get the value of t.

Example 12.7

The black buck population is growing in the Abohar (Punjab) sanctuary. An ornithologist is interested in estimating this population. For this purpose, he wishes to work out optimal sizes of first and second samples so that the survey to be conducted could yield reasonably accurate estimate. The costs per animal of catching-marking and catching-examining are Rs 100 and Rs 30 respectively. Work out optimal values of t and n if the total funds at disposal are Rs 30,000 and the overhead cost is Rs 2800. Based on the information provided by the residents of the area, the rough guess about black buck population is 3000.

Solution

We have C = 30,000, c_0 = 2800, c_1 = 100, c_2 = 30, and rough guess of N = 3000. Let us work out the values of K_1 and K_2 using (12.22). Therefore,

$$K_1 = \frac{Nc_1 - C + c_0}{c_1 - c_2}$$

$$= \frac{3000\,(100) - 30{,}000 + 2800}{100 - 30}$$

$$= 3897.14$$

$$K_2 = \left[\frac{c_1(Nc_2 - C + c_0)}{c_2\,(Nc_1 - C + c_0)}\right]^{\frac{1}{2}}$$

$$= \left[\frac{100\,\{3000(30) - 30{,}000 + 2800\}}{30\,\{3000\,(100) - 30{,}000 + 2800\}}\right]^{\frac{1}{2}}$$

$$= .876$$

Value of n is then obtained by using (12.21). This yields

$$n = K_1 \pm K_1 K_2$$

$$= 3897.14 \pm (3897.14)\,(.876)$$

$$= 483.25,\ 7311.03$$

Since n can not be greater than N, obviously, the admissible value of n is

$$n = 483.25 \approx 483$$

Now we work out optimum t by using (12.20). We thus have

$$t = \frac{C - c_0 - c_2 n}{c_1}$$

$$= \frac{30{,}000 - 2800 - 30(483.25)}{100}$$

$$= 127.02$$

$$\approx 127$$

Therefore, for the given budget and costs, the optimal values of t and n are 127 and 483 respectively. ■

12.4.2 *Optimum Sample Sizes for Minimum Cost*

If the funds are not restricted then for given α and A, the experimenter would wish to choose t and n so that the total cost is minimum. Given c_1, c_2, and rough guess of N, the investigator would choose the pair (t, n) satisfying the equations (12.17) and (12.19). The value of D for given α and A will be used from table 12.1. The solution to equations (12.17) and (12.19) provides optimal t and n.

Optimal t and n for given α, A, and rough guess of N :

$$t = \frac{Nc_2\,\delta(1-\delta)}{c_1 - c_2\delta^2} \tag{12.23}$$

$$n = \frac{N\delta(c_1 - c_2\delta)}{c_1 - c_2\delta^2} \tag{12.24}$$

where

$$\delta = \left[\frac{c_1 D}{c_2(N-1)}\right]^{\frac{1}{2}} \tag{12.25}$$

The appropriate value of D is taken from table 12.1.

Example 12.8

Suppose there is no fund restriction in example 12.7 .Then work out optimal sizes for the first and second samples, taking α = .05 and A = .25.

Solution

In this problem, we are to work out t and n for given α = .05, A = .25, c_1 = 100, and c_2 = 30. From table 12.1, for 1-α = .95 and A = .25, one can see that D = 69.9. Then from (12.25),

$$\delta = \left[\frac{c_1 D}{c_2(N-1)}\right]^{\frac{1}{2}}$$

$$= \left[\frac{(100)\,(69.9)}{(30)\,(3000-1)}\right]^{\frac{1}{2}}$$

$$= .2787$$

The use of (12.23) and (12.24) then yields the values of t and n as

$$t = \frac{Nc_2\delta(1-\delta)}{c_1 - c_2\delta^2}$$

$$= \frac{(3000)\,(30)\,(.2787)\,(1-.2787)}{100-(30)\,(.2787)^2}$$

$$= 185.24$$

$$\approx 185$$

$$n = \frac{N\delta(c_1 - c_2\delta)}{c_1 - c_2\delta^2}$$

$$= \frac{(3000)\,(.2787)\,[100-(30)\,(.2787)]}{100-(30)\,(.2787)^2}$$

$$= 784.47$$

$$\approx 784$$

For the given situation, the optimal sizes of the first and second samples are 185 and 784 respectively. ■

12.5 SOME FURTHER REMARKS

12.1 All the results from (12.16) to (12.25), as stated earlier, rest on the assumption of $N > 100$. In case $N < 100$, the reader should refer to Lieberman and Owen (1961) for methods that determine t and n optimally.

12.2 The procedure of single mark in closed population was extended to *multiple markings* by Schnabel (1938). A series of s samples of sizes $n_1, n_2, \ldots, n_s$ are drawn. Each sample captured, commencing from the second, is examined for marked members and then every member of the sample is given another mark before the sample is returned to the population. If marking/tagging is differentiated for different samples, then the capture-recapture history of any animal caught during the experiment can be inferred. A variety of Schnabel-type models are available. Some models are based on fixed sample sizes, and in others, the sample sizes are random variable. For details, the reader is referred to Seber (1973).

12.3 The estimation of population size in which there is possibly death, recruitment, immigration, and permanent emigration has been dealt with by Jolly (1965) and Seber (1965). Here, sequence of s samples of sizes $n_1, n_2, \ldots, n_s$ are drawn. For each sample, only m_i of n_i individuals are marked and returned to the population. This model also allows for accidental deaths due to marking and handling.

12.4 Pollock *et al.* (1984) have built a model that relates capture probabilities to auxiliary variables. The auxiliary variables may be of two types : (1) environmental variables which include air temperature, water temperature, humidity, and amount of effort expended in obtaining the sample; and (2) animal variables which affect an individual animal's probability of capture, and it includes age, weight, and body length.

12.5 Badaloni (1993) reviews a number of sampling techniques for the statistical analysis of sparse biological populations. In particular, he focuses on the line transect and the line intersection methods. These techniques are widely used in the natural sciences for estimating the population density.

LET US DO

12.1 The capture and recapture method for estimating mobile populations rests on certain assumptions. What are these ?

12.2 In what sense the Bailey's estimator $\hat{N}_b$ in (12.8) is superior to Petersen estimator $\hat{N}_p$ given in (12.6) ?

12.3 Describe the negative hypergeometric model for estimating mobile populations.

12.4 How does the direct sampling differ from inverse sampling ? Discuss with special reference to wildlife populations.

12.5 Describe the procedure for arriving at the optimal sizes of first and second samples in case the funds at disposal are limited.

12.6 How should the investigator determine the optimal values of t and n for given α and A, so that, the total cost is minimum ?

12.7 A survey was undertaken to estimate frog population living in a certain pond. Frogs generally live about the margins of permanent and semi-permanent areas of water. Seventy frogs were taken from random points around the circumference of the pond. The captured frogs were marked by toe-clipping and released into the population. After allowing sufficient time to mix with the unmarked frogs, a WOR random sample of 150 frogs was drawn. In this sample, 30 frogs were found marked. Estimate the frog population inhabiting the pond, and obtain confidence limits for it.

12.8 From the data given in example 12.4, estimate the total underground ant population using estimator $\hat{N}_p$ in (12.6), and also work out confidence interval for it.

12.9 The pigeons are causing severe loss to the grains stored in a warehouse scattered over an area of 4 hectares. An ornithologist wishes to estimate the pigeon population inhabiting this area. Ninety six pigeons were trapped, banded, and then released into the population. On the next day, 287 pigeons were observed with binoculars. Out of these, 19 pigeons were found banded. Assuming the population to be closed, estimate the total number of pigeons in this area using Bailey's estimator $\hat{N}_b$ in (12.8). Also, determine confidence interval for it.

12.10 For t = 27, n = 23, and s = 17 in example 12.2, estimate the Snowshoe Hares population using Bailey's estimator in (12.8). Also, determine the bounds on the error of estimation.

12.11 In order to estimate the size of the brook trout (*Salvelinus fontinalis*) population in a pond, 116 brook trouts were captured with trap nets, and released into the pond with jaw-tags. After allowing sufficient time for the marked fish to mingle with the unmarked ones, it was decided to capture 26 marked trouts. To get 26

marked trouts using WOR random sampling, 289 trouts had to be caught. Using estimator $\hat{N}_{ib}$ given in (12.10), estimate total number of brook trouts in the pond, and also work out confidence interval for N.

12.12 The Department of Zoology undertook a survey to estimate the black buck population in the Abohar sanctuary. Using optimal t and n from example 12.8, 185 black bucks were trapped from random places in the sanctuary and released after tagging. The tags used were the leather collars studded with shinning colored plastic buttons. After 10 days, the area was surveyed again with binoculars, and the black bucks observed for the collars. In all, 784 bucks were observed and 40 of them found tagged. Estimate the number of black bucks in the sanctuary, and also find confidence interval for the population size N.

12.13 Four years ago, the estimated population of rabbits in a preserve was 15,000. The wildlife biologist feels that the rabbit population has declined due to hunting by the nearby residents. The rough guess of the present population is around 13,000. The biologist wishes to estimate the population with reasonable degree of accuracy. Hence, he needs optimal sizes for first and second samples. The costs per rabbit for catching-marking and observing are Rs 15 and 6 respectively. Using $\alpha = .05$ and $A = .10$, determine the optimal values for t and n.

12.14 For the data of exercise 12.13, estimate optimal t and n if the total funds allotted for the purpose are Rs 15,000 with Rs 1,500 as the overhead cost, so that, the resulting estimate has a maximum possible accuracy.

CHAPTER 13

Nonresponse Errors

13.1 INTRODUCTION

In the preceding chapters, several survey designs have been discussed at length with respect to their applications. For each design, it was assumed that the true values of the variables of interest could be made available for the elements of the population under consideration. However, this is not usually the case in practice. The errors can occur at almost every stage of planning and execution of survey. These errors may be attributed to various causes right from the beginning stage, where the survey is planned and designed, to the final stage when the data are processed and analyzed. This gives rise to the following definition of nonsampling errors.

Definition 13.1 The errors arising mainly due to misleading definitions and concepts, inadequate frames, unsatisfactory questionnaire, defective methods of data collection, tabulation, coding, decoding, incomplete coverage of sample units, etc., are called *nonsampling errors.*

The incomplete coverage of units, mentioned in definition 13.1 above, occurs due to nonavailability of information from some units included in the sample. It happens if a questionnaire is mailed to a sample of units, and some respondents fail to return the completed questionnaire. If the visits are made to a sample of households, some respondents may be away from home, and others may refuse to co-operate. This leads us to definition 13.2.

Definition 13.2 The inability to collect relevant information for some of the sample units due to refusal by respondents to divulge information, their being not-at-home, sample units being inaccessible, or due to any other such reason, is termed nonresponse. The errors resulting from this incomplete coverage of the sample are called *nonresponse errors.*

Nonresponse is one of the several kinds of errors included in nonsampling errors. In this chapter, we shall only consider the problem of nonresponse. For details about other kinds of errors included in nonsampling errors, reader may refer to Sukhatme *et al.* (1984), Cochran (1977), and Zarkovich (1966).

When the respondents do not send back the required information with respect to the questionnaire mailed by the investigator, the available sample of returns is

incomplete. The resulting nonresponse is sometimes so large that it vitiates the results. While nonresponse can not be completely eliminated in practice, it could be overcome to a great extent by persuasion, or by some other methods. One way of dealing with this problem is due to Hansen and Hurwitz (1946).

13.2 HANSEN AND HURWITZ TECHNIQUE

Suppose that N units of the population can be divided into two classes. Population units (respondents) that will return the completed questionnaire, mailed to them by the investigator, without any reminders being sent, shall constitute the response class, whereas the remaining population units shall belong to the nonresponse class. Let N_1 and N_2 be the number of units in the population that form the response and nonresponse classes respectively, so that, $N = N_1+N_2$. Further, suppose that out of n units selected with equal probabilities and WOR sampling, n_1 units respond and n_2 $(= n-n_1)$ units do not respond in the first attempt. Let a WOR random subsample of h_2 units be selected from the n_2 nonresponding units, such that $n_2 = h_2 f$, for collecting information by special efforts like repeated reminders or personal interviews. Then define

$$\bar{G}_w = \frac{1}{n}\left(\sum_{i=1}^{n_1} y_i^2 + \frac{n_2}{h_2}\sum_{i=1}^{h_2} y_i^2\right) \tag{13.1}$$

$$s_{h2}^2 = \frac{1}{h_2-1}\left(\sum_{i=1}^{h_2} y_i^2 - h_2\bar{y}_{h2}^2\right) \tag{13.2}$$

where

$$\bar{y}_{h2} = \frac{1}{h_2}\sum_{i=1}^{h_2} y_i$$

We now present the following results :

Unbiased estimator of population mean $\bar{Y}$:

$$\bar{y}_w = \frac{1}{n}(n_1\bar{y}_1 + n_2\bar{y}_{h2}) \tag{13.3}$$

where $\bar{y}_1$ is the sample mean based on n_1 units, and $\bar{y}_{h2}$ is defined in (13.2) above.

Variance of estimator $\bar{y}_w$:

$$V(\bar{y}_w) = \left(\frac{1}{n}-\frac{1}{N}\right)S^2 + \frac{f-1}{n}\left(\frac{N_2}{N}\right)S_2^2 \tag{13.4}$$

where S^2 and S_2^2 are the mean squares for the whole population and the nonresponse class in the population respectively.

Estimator of variance $V(\bar{y}_w)$:

$$v(\bar{y}_w) = \frac{n(N-1)}{N(n-1)}\left[\frac{N-n}{n(N-1)}(\bar{G}_w - \bar{y}_w^2) + \frac{f-1}{n}\left(\frac{n_2}{n}\right)s_{h2}^2\right] \tag{13.5}$$

The second term in (13.4) vanishes for f = 1, which should be the case if it is possible to elicit information from each of the n_2 nonrespondents.

Example 13.1

There are 570 progressive farmers in a state. The investigator is interested in estimating the average cost of herbicides used per hectare, by progressive farmers, for the paddy crop. A WOR simple random sample of 40 farmers was selected. The survey questionnaire was mailed to the selected farmers. Only 16 farmers returned the completed questionnaire. Out of the remaining 24 farmers, a further subsample of 7 farmers was selected. These selected farmers were personally contacted, and the required information obtained. The data collected are given in table 13.1.

Table 13.1 Cost (in rupees) of herbicides used per hectare

Through mail				Through personal interview	
Farmer	Cost of herbicides	Farmer	Cost of herbicides	Farmer	Cost of herbicides
1	185	9	150	1	196
2	170	10	181	2	147
3	191	11	167	3	213
4	211	12	195	4	189
5	160	13	174	5	192
6	250	14	197	6	176
7	176	15	213	7	200
8	182	16	186		

Estimate the average cost of herbicides used per hectare by the progressive farmers of the state, and work out the confidence interval for it.

Solution

From the statement of the example, we have N = 570, n = 40, n_1 = 16, n_2 = 24, and h_2 = 7. First of all, we compute means $\bar{y}_1$ and $\bar{y}_{h2}$ based on n_1 and h_2 observations respectively. Thus, we have

$$\bar{y}_1 = \frac{1}{16}(185+170+\ldots+186)$$

$$= \frac{2988}{16} \qquad \text{[from columns (2) and (4) of table 13.1]}$$

$$= 186.75$$

$$\bar{y}_{h2} = \frac{1}{7}(196+147+...+200)$$

$$= \frac{1313}{7} \qquad \text{[from column (6) of table 13.1]}$$

$$= 187.57$$

The estimate of the average cost of herbicides used per hectare is then obtained by using (13.3). This is

$$\bar{y}_w = \frac{1}{n}(n_1\bar{y}_1 + n_2\bar{y}_{h2})$$

$$= \frac{1}{40}[(16)(186.75) + (24)(187.57)]$$

$$= 187.24$$

For obtaining the estimate of variance, the expression for $\bar{G}_w$ in (13.1), and that for s^2_{h2} in (13.2), are to be evaluated. We, therefore, have

$$\bar{G}_w = \frac{1}{n}\left(\sum_{i=1}^{n_1} y_i^2 + \frac{n_2}{h_2}\sum_{i=1}^{h_2} y_i^2\right)$$

$$= \frac{1}{40}\left[(185)^2 + (170)^2 + ... + (186)^2 + \frac{24}{7}\{(196)^2 + (147)^2 + ... + (200)^2\}\right]$$

$$= \frac{1}{40}\left[566552 + \frac{24}{7}(248955)\right]$$

$$= 35502.80$$

$$s^2_{h2} = \frac{1}{h_2 - 1}\left(\sum_{i=1}^{h_2} y_i^2 - h_2\bar{y}^2_{h2}\right)$$

$$= \frac{1}{7-1}[248955 - 7(187.57)^2]$$

$$= 446.24$$

Now the variance estimator in (13.5) is given by

$$v(\bar{y}_w) = \frac{n(N-1)}{N(n-1)}\left[\frac{N-n}{n(N-1)}(\bar{G}_w - \bar{y}_w^2) + \frac{f-1}{n}\left(\frac{n_2}{n}\right)s^2_{h2}\right]$$

Substituting the values of different terms, one obtains

$$v(\bar{y}_w) = \frac{40(570-1)}{570(40-1)}\left[\frac{(570-40)}{40(570-1)}\{35502.80 - (187.24)^2\}\right.$$

$$\left. + \left\{\frac{(24/7)-1}{40}\right\}\left(\frac{24}{40}\right)(446.24)\right]$$

$$= (1.0238)[(.0233)(443.982) + (.0607)(.6)(446.24)]$$

$$= 27.2298$$

The required confidence interval is obtained by using the formula

$$\bar{y}_w \pm 2\sqrt{v(\bar{y}_w)}$$

On substituting for different terms, it reduces to

$$\begin{aligned} & 187.24 \pm 2\sqrt{27.2298} \\ = \; & 187.24 \pm 10.44 \\ = \; & 176.80,\ 197.68 \end{aligned}$$

This indicates that had all the 570 farmers been surveyed, the average cost of herbicides used per hectare would have taken a value in the closed interval [176.80, 197.68] with approximate probability .95. ■

A *cost function* appropriate for the Hansen and Hurwitz technique is given by

$$C' = c_0 n + c_1 n_1 + c_2 h_2 \tag{13.6}$$

where

c_0 = the cost of including a sample unit in the initial survey,
c_1 = the per unit cost of collecting, editing, and processing information on the study variable in the response class, and
c_2 = the per unit cost of interviewing and processing same information in the nonresponse class.

As C' varies from sample to sample, we consider expected cost

$$E(C') = C = \frac{n}{N}\left(Nc_0 + N_1 c_1 + \frac{N_2}{f} c_2\right) \tag{13.7}$$

We now determine the optimum values of n and f for which C is minimum and the variance $V(\bar{y}_w) = V_0$.

Optimum n and f for fixed variance V_0 that minimizes cost in (13.7) :

$$f = \left[\frac{c_2(S^2 - N_2 S_2^2/N)}{S_2^2 (c_0 + N_1 c_1/N)}\right]^{\frac{1}{2}} \tag{13.8}$$

$$n = \frac{S^2 + \{N_2(f-1)\, S_2^2\}/N}{V_0 + S^2/N} \tag{13.9}$$

where the symbols involved have already been defined.

Example 13.2

In a survey, the expected response rate is 40 percent, $S_2^2 = (3/4)S^2$, it costs \$ .5 to include a unit in the sample, \$2 per unit to observe study variable in the response class, and \$6 per unit to observe the study variable in the nonresponse class. Determine optimum

values of f and n if the population mean is to be estimated with a tolerable variance $V_0 = S^2/200$. Also, work out the total expected cost of the survey with the cost function considered in (13.7).

Solution

The statement of the example provides

$$N_1 = N\,\frac{40}{100},\quad N_2 = N\,\frac{60}{100},\quad S_2^2 = \frac{3}{4}\,S^2,\quad c_0 = .5,\ c_1 = 2,\ \text{and}\ c_2 = 6.$$

On making substitutions in (13.8), one arrives at the optimum value of f as

$$f = \left[\frac{6\{S^2 - (60)\,(3S^2)/400\}}{(3S^2/4)\,\{.5 + (40)\,(2)/100\}}\right]^{\frac{1}{2}}$$

$$= \left[\frac{6\{1 - (60)\,(3)/400\}}{(3/4)\,\{.5 + (40)\,(2)/100\}}\right]^{\frac{1}{2}}$$

$$= 1.84$$

The optimum value of n is given by (13.9). On substituting different values, one gets

$$n = \frac{S^2 + (60)\,(1.84 - 1)\,(3S^2)/400}{(S^2/200) + S^2/N}$$

$$= \frac{200\,[1 + (60)\,(1.84 - 1)\,(3)/400]}{1 + 200/N}$$

For large N, it yields

$$n = 275.6 \approx 276.$$

From (13.7), the total expected cost (in dollars) of the survey, is found as

$$C = 276\left[.5 + \frac{40}{100}(2) + \left(\frac{60}{100}\right)\left(\frac{1}{1.84}\right)(6)\right]$$

$$= 898.79\ \blacksquare$$

Hansen and Hurwitz technique loses its merit when nonresponse is large. Durbin (1954) observed that when $S_2^2 = S^2$, and cost of collecting data in the nonresponse class is much higher than that in the response class, it will not be worthwhile to go for this technique.

El-Badry (1956) extended Hansen and Hurwtiz technique further. He suggested to send repeated waves of questionnaires to the nonresponding units. As soon as the investigator feels that further waves will not be much effective, a subsample from the remaining nonresponding units is selected and the information is collected through personal interview. The required estimate is based on the total information collected from all the attempts.

Deming (1953) has advocated the use of *call-back technique*. According to him, if the chosen sample member is not at home or is unable to take part in an interview at the time of call, then it is advisable for the interviewer to call back. He has shown how successive recalls help in reducing the bias, and proposed a method to arrive at optimum number of recalls to achieve a given precision. He has also determined the optimum number of call-backs for the fixed total cost of survey by minimizing the variance.

An interesting plan dealing with the reduction of bias *without call-backs* is also available. This is presented in the following section.

13.3 BIAS REDUCTION WITHOUT CALL - BACKS

An ingenious technique of reducing biases present in the results of the first call was given by *Politz and Simmons* (1949), for surveys where information on study variable is collected through interview method. The proposed procedure attempts to reduce the bias, owing to incomplete sample, without resorting to successive call-backs. According to this technique, the calls are made during the evening on the six weekdays. The time of the call is assumed random within the interview hours. If the respondent is available at home, the required information is obtained. Besides, he is also asked how many times he was at home, at the time of visit, on each of the five preceding weekdays ? Suppose that n respondents were selected by using WR equal probability sampling. Further, let p_i be the probability that the i-th respondent is found at home at the time of call. Then, estimate of p_i is given by

$$\hat{p}_i = \frac{t+1}{6}, \quad t = 0, 1, \ldots, 5 \tag{13.10}$$

where t is the number of times the respondent was at home, at the time of call, during the last five evenings. Defining

$$Q_i = \sum_{t=0}^{5} \frac{1}{(t+1)} \binom{6}{t+1} p_i^{t+1} q_i^{5-t} \tag{13.11}$$

where $q_i = 1-p_i$, the estimator of population mean and the other related results, for this procedure, are then as given in (13.12) through (13.15).

Estimator of population mean $\bar{Y}$:

$$\bar{y}_{ps} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\hat{p}_i} \tag{13.12}$$

where y_i assumes value zero if the respondent is not found at home at the time of call.

Bias of estimator $\bar{y}_{ps}$:

$$B(\bar{y}_{ps}) = -\frac{1}{N} \sum_{i=1}^{N} Y_i q_i^6 \tag{13.13}$$

Variance of estimator $\bar{y}_{ps}$:

$$V(\bar{y}_{ps}) = \frac{1}{n}\left[\frac{6}{N}\sum_{i=1}^{N} Y_i^2 Q_i - \left\{\frac{1}{N}\sum_{i=1}^{N} Y_i\,(1-q_i^6)\right\}^2\right] \tag{13.14}$$

where Q_i has been defined in (13.11).

Estimator of variance $V(\bar{y}_{ps})$:

$$\begin{aligned} v(\bar{y}_{ps}) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{\hat{p}_i} - \bar{y}_{ps}\right)^2 \\ &= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{y_i}{\hat{p}_i}\right)^2 - n\bar{y}_{ps}^2\right] \end{aligned} \tag{13.15}$$

Example 13.3

The Wipro company is engaged in marketing of personal computers (PC). For taking certain decisions, the company needs information on the average annual maintenance cost for a personal computer purchased three to five years ago. For this purpose, a WR random sample of 28 buyers was selected from 640 buyers, to whom the company had sold personal computers during the last three to five years. The buyers were interviewed in the afternoon of the weekdays. The information in respect of annual maintenance cost (y) in rupees, and the number of days on which the respondent was available at the time of call during the last five days (denoted by t), is exhibited in table 13.2. The estimation variable assumes zero value if the respondent is not available at the time of call. Marks "—" against respondents at serial numbers 9 and 20 indicate their nonavailability.

Table 13.2 The data collected from the sampled buyers

Respondent	y_i	t	$\hat{p}_i = \frac{t+1}{6}$	Respondent	y_i	t	$\hat{p}_i = \frac{t+1}{6}$
1	500	3	.6667	15	900	5	1.0000
2	1000	5	1.0000	16	1300	4	.8333
3	3500	4	.8333	17	560	5	1.0000
4	400	4	.8333	18	700	3	.6667
5	2600	5	1.0000	19	0	3	.6667
6	1500	3	.6667	20	—	—	—
7	800	5	1.0000	21	1600	5	1.0000
8	4100	3	.6667	22	1000	4	.8333
9	—	—	—	23	520	4	.8333
10	710	3	.6667	24	1300	5	1.0000
11	2100	4	.8333	25	2500	3	.6667
12	1050	4	.8333	26	0	2	.5000
13	0	2	.5000	27	1200	4	.8333
14	1125	3	.6667	28	300	1	.3333

Estimate the average annual repair charges per PC, and obtain confidence interval for it.

Solution

From the statement of the example, we have N = 640 and n = 28. The estimate of average annual maintenance cost per PC is computed by using (13.12). The formula is

$$\bar{y}_{ps} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\hat{p}_i}$$

The estimates $\hat{p}_i$ are computed and given in table 13.2. Making substitutions for n, y_i, and $\hat{p}_i$, yields

$$\bar{y}_{ps} = \frac{1}{28}\left(\frac{500}{.6667} + \frac{1000}{1.0000} + \ldots + \frac{300}{.3333}\right)$$

$$= \frac{1}{28}(39646.28)$$

$$= 1415.94$$

as an estimate of average annual maintenance cost.

The variance estimator is provided by (13.15). We, therefore, write

$$v(\bar{y}_{ps}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{\hat{p}_i} - \bar{y}_{ps}\right)^2$$

$$= \frac{1}{28(28-1)}\left[\left(\frac{500}{.6667} - 1415.94\right)^2 + \left(\frac{1000}{1.0000} - 1415.94\right)^2 + \ldots + \left(\frac{300}{.3333} - 1415.94\right)^2\right]$$

$$= \frac{1}{28(28-1)}\left[\left(\frac{500}{.6667}\right)^2 + \left(\frac{1000}{1.0000}\right)^2 + \ldots + \left(\frac{300}{.3333}\right)^2 - 28(1415.94)^2\right]$$

$$= \frac{1}{28(28-1)}[109319425 - 56136808]$$

$$= 70347.374$$

The confidence limits will be obtained by using

$$\bar{y}_{ps} \pm 2\sqrt{v(\bar{y}_{ps})}$$

$$= 1415.94 \pm 2\sqrt{70347.374}$$

$$= 1415.94 \pm 530.46$$

$$= 885.48,\ 1946.40$$

Above confidence limits indicate that the average annual maintenance cost per PC, for the population, is covered by the interval [885.48, 1946.40] with approximate probability .95. ■

13.4 WARNER'S RANDOMIZED RESPONSE MODEL

Sample surveys on human populations have established that the innocuous questions usually receive good response, whereas questions on sensitive items involving controversial assertions, stigmatizing and/or incriminating matters, which people like to hide from others, excite resistance. Direct questions about them often result in either refusal to respond, or falsification of their answers. This introduces *nonresponse error* which makes the estimation of relevant parameters, for instance, proportion of population belonging to sensitive group, unreliable. Recognizing the fact that such biases are frequent when respondents are queried directly about sensitive or embarrassing matters, Warner (1965) developed an ingenious interviewing procedure to reduce or eliminate these evasive answer biases.

This model was built on the hope that the co-operation might be better when the respondent is requested for information in anonymity, provided by randomization device, rather than on direct question basis. The randomization device creates a stochastic relationship between the question and the individual's response, and thus provides protection to confidentiality of the respondent. The information from the respondents is elicited in terms of "yes" or "no" answers without endangering their privacy. He called the procedure as *randomized response* (RR) *technique*. We first discuss the Warner's (1965) pioneer model, and then some other popular modifications and improvements.

The procedure involves the use of two statements, each of which divides the population into two mutually exclusive and complementary classes, say, A and not-A. In order to estimate π, the proportion of respondents in sensitive group A, every person selected in a WR simple random sample of n respondents is given a suitable randomization device consisting of two mutually exclusive statements of the form :

1. "I belong to sensitive group A"
2. "I belong to group not-A".

The statements (1) and (2) are represented in the randomization device with probability p and (1-p) respectively. The randomization device, for instance, may be a deck of cards of which a proportion p carries the first statement while on the remaining cards second statement is written. It could also be a spinner in which a sector formed by an angle of 360p degrees is marked with first statement, and the remaining 360(1-p) degrees sector is marked with second statement. The respondent is asked to choose one of the two statements randomly and to answer "yes" if the selected statement points to his group with respect to the attribute A, "no" otherwise. The interviewee, however, does not reveal to the interviewer as to which of the two statements has been chosen. It is assumed that these "yes" and "no" answers are reported truthfully. The formulas for estimator of proportion, its variance, and estimator of variance are given in (13.16) to (13.18).

Unbiased estimator of proportion π :

$$\hat{\pi}_w = \frac{(n'/n) - 1 + p}{2p - 1}, \quad p \neq .5, \tag{13.16}$$

where n' is the number of persons answering "yes".

Variance of estimator $\hat{\pi}_w$:

$$V(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2} \tag{13.17}$$

Unbiased estimator of variance $V(\hat{\pi}_w)$:

$$v(\hat{\pi}_w) = \frac{(n'/n)(1 - n'/n)}{(n-1)(2p-1)^2} \tag{13.18}$$

The expression (13.17) reveals that it is desirable to choose p as close to 1 or 0 as possible without threatening the degree of co-operation by the respondent.

The RR technique was extended to polychotomous populations by Abul-Ela *et al.* (1967). Some test procedures for detecting untruthful answering by the respondents, for Warner's (1965) model, have been developed by Lakshmi and Raghavarao (1992), and Krishnamoorthy and Raghavarao (1993).

Example 13.4

The Mathura depot of State Transport Corporation has received complaints that some of the bus drivers resort to drunken driving. The General Manager of the depot has been asked by the state administration to start a campaign creating awareness against drunken driving among the drivers. The General Manager thought it wise to estimate the proportion of such drivers. In order to accomplish the objective, a WR simple random sample of 120 drivers, out of a total of 908 drivers, was selected. Because of the sensitive nature of inquiry, it was decided to use randomized response technique. The randomization device used consisted of a deck of 100 cards, 80 of which bore the statement "I have resorted to drunken driving". The remaining 20 cards carried the statement "I have not resorted to drunken driving".

The interviewee was asked to select a card randomly from the deck after shuffling it, and report "yes" if the statement written on the card pointed to his actual status, and "no" otherwise. He was not to tell the interviewer as to which of the two statements had been chosen. In all, there were 36 "yes" and 84 "no" responses. Using this information, estimate the proportion of bus drivers who resort to drunken driving. Also, obtain confidence limits for it.

Solution

In this case, we have n = 120, n' = 36, and p = 80/100 = .8. The required estimate is obtained by using (13.16). Therefore,

$$\hat{\pi}_w = \frac{(n'/n) - 1 + p}{2p - 1}$$

$$= \frac{(36/120) - 1 + .8}{2\ (.8) - 1}$$

$$= .167$$

We now work out the estimate of variance from (13.18). Thus, we have

$$v(\hat{\pi}_w) = \frac{(n'/n)\ (1 - n'/n)}{(n-1)\ (2p-1)^2}$$

On making substitutions, one gets

$$v(\hat{\pi}_w) = \frac{(36/120)[1 - (36/120)]}{(120-1)\ [2\ (.8) - 1]^2}$$

$$= .004902$$

The confidence limits for π are

$$\hat{\pi}_w \pm 2\sqrt{v(\hat{\pi}_w)}$$

$$= .167 \pm 2\sqrt{.004902}$$

$$= .167 \pm .140$$

$$= .027, .307$$

Thus, from the sample data, the proportion of drivers who had resorted to drunken driving is estimated as .167. Also, its population value is most likely to fall in the confidence interval [2.7, 30.7] percent. ■

The large width of the confidence interval is due to large variance. The variance of the estimator $\hat{\pi}_w$ gets inflated as it contains an additional component of variance because of the randomization of response. The method, therefore, generally requires a large sample size to obtain a reasonably small variance of the estimator.

In the Warner's model, the sample size is fixed and the number of "yes" answers is a random variable. Mangat and Singh (1991a) have given a procedure in which the number of "yes" answers is fixed but the sample size is a random variable. They have given another modification known as two-stage procedure. This technique is discussed in the next section.

13.5 MANGAT AND SINGH'S TWO-STAGE MODEL

In Mangat and Singh's (1990) method, each interviewee in the WR simple random sample of n respondents is provided with two randomization devices R_1 and R_2. The randomization device R_1 consists of two statements, namely :

1. " I belong to sensitive group A"
2. " Go to randomization device R_2",

represented with probabilities T and (1-T) respectively. The randomization device R_2 which uses two statements,

1. "I belong to sensitive group A"
2. "I belong to group not-A",

with known probabilities p and (1-p), is exactly the same as used by Warner (1965). The interviewee is instructed to experience first the randomization device R_1. One is to use R_2 only if directed by the outcome of R_1. The respondent is required to answer "yes" if the outcome points to the attribute he/she possesses, and answer "no" if the complement of his status is pointed out by the outcome. The estimator and the related results for this model are given below :

Unbiased estimator of population proportion π :

$$\hat{\pi}_{ms} = \frac{n'/n - (1-T)(1-p)}{2p-1+2T(1-p)} \tag{13.19}$$

where, as in Warner's model, n' is the number of "yes" answers.

Variance of estimator $\hat{\pi}_{ms}$:

$$V(\hat{\pi}_{ms}) = \frac{\pi(1-\pi)}{n} + \frac{(1-T)(1-p)[1-(1-T)(1-p)]}{n[2p-1+2T(1-p)]^2} \tag{13.20}$$

Estimator of variance $V(\hat{\pi}_{ms})$:

$$v(\hat{\pi}_{ms}) = \frac{(n'/n)(1-n'/n)}{(n-1)[2p-1+2T(1-p)]^2} \tag{13.21}$$

Mangat and Singh's strategy can always be made more efficient than the Warner's model by suitably choosing the value of T for any practical value p. Such a value of T is given by

$$T > \frac{1-2p}{1-p} \tag{13.22}$$

The estimator $\hat{\pi}_{ms}$ using a value of T satisfying above inequality is, therefore, expected to yield a confidence interval of smaller width than the one provided by Warner's model. However, if the probability for drawing the statement on sensitive character in Warner's model is taken equal to T+p(1-T), and it is retained at level p in the second randomization device of Mangat and Singh's model, then both the strategies are equally efficient.

Example 13.5

For example 13.4, estimate the proportion of drivers resorting to drunken driving, using Mangat and Singh's two-stage procedure, assuming that the number of "yes" answers received was 30. Take the value of T as .3.

Solution

Now we have $n = 120$, $n' = 30$, $p = .8$, and $T = .3$. The estimate of the required proportion is worked out by using (13.19). Thus,

$$\hat{\pi}_{ms} = \frac{(30/120) - (1-.3)(1-.8)}{2(.8) - 1 + 2(.3)(1-.8)}$$

$$= .153$$

The estimator of variance in this case is provided by (13.21). Hence,

$$v(\hat{\pi}_{ms}) = \frac{(n'/n)(1-n'/n)}{(n-1)[2p-1+2T(1-p)]^2}$$

$$= \frac{(.25)(1-.25)}{(120-1)[2(.8)-1+2(.3)(1-.8)]^2}$$

$$= .003039$$

The confidence limits for the population proportion are, therefore, calculated as

$$\hat{\pi}_{ms} \pm 2\sqrt{v(\hat{\pi}_{ms})}$$

$$= .153 \pm 2\sqrt{.003039}$$

$$= .153 \pm .110$$

$$= .043, .263$$

The estimate of the proportion of bus drivers who resort to drunken driving is thus found to be .153. Also, the confidence limits indicate that the population proportion is likely to vary from 4.3% to 26.3%. ■

Mangat and Singh (1991b) have extended their above procedure to the situation where the respondents are selected using SRS without replacement method. For this case also, they considered the estimator in (13.19). However, the estimator of variance is little complicated for this procedure. Several other workers have also modified the Warner's (1965) model. Franklin (1989) and Singh and Singh (1992, 1993) use continuous randomization device instead of discrete one. Mangat (1994) and Mangat *et al.* (1995) have proposed some other variants of Warner's model.

Walt R. Simmons felt that the confidence of the respondents might be enhanced further if one of the two questions refers to a nonsensitive attribute (say) y, unrelated to the stigmatized characteristic. Following his suggestion, Horvitz *et al.* (1967) developed a procedure, and called it *unrelated question randomized response model.* In short, we shall call it "U-model".

13.6 UNRELATED QUESTION MODEL

Though the *unrelated question model* was proposed by Horvitz *et al.* (1967), but the theoretical framework for it was developed by Greenberg *et al.* (1969). While developing theory, they dealt with both the situations where π_y, the proportion of innocuous character in the population, is known and when it is unknown.

13.6.1 *Case I - π_y Unknown*

For the situation when π_y is not known, two samples of sizes n_1 and n_2 respondents are drawn from the population using SRS with replacement, so that, $n_1 + n_2 = n$ is the overall sample size. Each sample is then used to collect information on both the characters. In this model, two sets of randomization device need to be used. Each device consists of the following two statements :

1. "I am a member of group A"
2. "I am a member of group Y",

where, as already mentioned, group A consists of units possessing the sensitive characteristic, and membership in group Y carries no embarrassment. Statement regarding character y in the randomization device could be -"Does your date of birth fall in the month of January or February ?". π_y, the probability for the statement to hold for any respondent, could be manipulated by increasing or decreasing the length of time period in which birth date is to fall. However, actual value of π_y may or may not be known. In the first randomization device, let group A be represented with probability p_1 and the group Y with $(1-p_1)$, whereas in the second device the groups A and Y be represented respectively with the probabilities p_2 and $(1-p_2)$. Let the first device be provided to each of the respondents in the sample of size n_1, and the second device is used for the respondents in the sample of size n_2. Each respondent is asked to select a statement randomly, and unobserved by the interviewer, from the device provided to him. He/she is required to report "yes" if the selected statement points to his/her actual status, and "no" otherwise. For i = 1, 2, let

n'_i = number of "yes" answers reported in the i-th sample, and

$$\theta_i = p_i\pi + (1-p_i)\,\pi_y \qquad (13.23)$$

be the probability that a "yes" answer will be reported by the respondents in the i-th sample. Postulating that the respondents report complete truth, the estimator and other related results are as follows :

Unbiased estimator of π when π_y is not known :

$$\hat{\pi}_g = \frac{1}{p_1 - p_2}\left[(1-p_2)\,\frac{n'_1}{n_1} - (1-p_1)\,\frac{n'_2}{n_2}\right] \qquad (13.24)$$

Variance of estimator $\hat{\pi}_g$:

$$V(\hat{\pi}_g) = \frac{1}{(p_1 - p_2)^2}\left[(1-p_2)^2\,\frac{\theta_1(1-\theta_1)}{n_1} + (1-p_1)^2\,\frac{\theta_2(1-\theta_2)}{n_2}\right] \qquad (13.25)$$

Estimator of variance $V(\hat{\pi}_g)$:

$$v(\hat{\pi}_g) = \frac{1}{(p_1 - p_2)^2}\left[(1-p_2)^2 \frac{(n_1'/n_1)(1-n'_1/n_1)}{n_1 - 1} + (1-p_1)^2 \frac{(n'_2/n_2)(1-n'_2/n_2)}{n_2 - 1}\right] \quad (13.26)$$

In order to minimize the variance of the estimator $\hat{\pi}_g$, Greenberg *et al.* (1969) deduced the following rules :

1. Choose one of the p_i, i = 1, 2, as close to 1 and the other as close to zero as the respondents are likely to accept, such that $p_1 + p_2 = 1$.
2. Choose π_y close to 0 or 1 according as $\pi < .5$ or $\pi > .5$. If $\pi = .5$, then $|\pi_y - .5|$ could be maximum on either side. While choosing π_y close to 0 or 1, depending on π, the investigator should take care that it should not be selected too close to 0 or 1, as it may affect the likelihood of co-operation by the respondents, and thus contradict the whole purpose of using unrelated question approach.
3. In practical situations, the total sample size $n(= n_1 + n_2)$ is fixed. One is then concerned with the choice of n_1 and n_2, so that the variance $V(\hat{\pi}_g)$ is minimized. The optimal allocation of n into n_1 and n_2 is given by

$$\frac{n_1}{n_2} = \frac{(1-p_2)\sqrt{\theta_1(1-\theta_1)}}{(1-p_1)\sqrt{\theta_2(1-\theta_2)}} \quad (13.27)$$

In application of rule (13.27), it is necessary to use a rough guess of π and π_y in order to calculate θ_1 and θ_2 defined in (13.23).

Example 13.6

A tough and neck to neck campaign is on between two candidates A and B in ward number 56 of a metropolitan city during elections for municipal corporation. Owing to surcharged and tense atmosphere, voters hesitate to divulge openly as to which candidate they would vote. In order to estimate the voting behavior in advance, two independent SRS with replacement samples - one consisting of 160 voters and other of 140 voters - were drawn from the frame (voters' list) consisting of 4160 voters.

Two decks of cards were prepared. The deck I consisted of two types of cards bearing statements :

1. "I shall vote for candidate A"
2. "I was born in the month of November",

occurring with probabilities .8 and .2 respectively. In deck II the statements (1) and (2) respectively were represented with probabilities .2 and .8 . The deck I was used for the voters in the first sample and deck II for the voters of the second sample. Each voter was asked to choose a statement randomly, unobserved by the interviewer, and say "yes"

if the selected statement points to his actual status, and "no" otherwise. On completion of survey, it was found that 55 "yes" answers had been reported by the respondents in the first sample, and 68 "yes" answers had come from the second sample.

Estimate the proportion of voters favoring candidate A, and also obtain confidence interval for it.

Solution

From the statement of the example we have $n_1 = 160$, $n_2 = 140$, so that, $n = 160 + 140 = 300$. Also, $p_1 = .8$, $p_2 = .2$, $n'_1 = 55$, and $n'_2 = 68$. The estimate of the proportion of voters who would vote for candidate A, is computed by using (13.24). Therefore,

$$\hat{\pi}_g = \frac{1}{p_1 - p_2}\left[(1-p_2)\frac{n'_1}{n_1} - (1-p_1)\frac{n'_2}{n_2}\right]$$

$$= \frac{1}{(.8-.2)}\left[(1-.2)\frac{55}{160} - (1-.8)\frac{68}{140}\right]$$

$$= .2964$$

The estimate of variance is provided by (13.26). The expression for this estimator is

$$v(\hat{\pi}_g) = \frac{1}{(p_1 - p_2)^2}\left[(1-p_2)^2 \frac{(n'_1/n_1)(1-n_1'/n_1)}{n_1 - 1} + (1-p_1)^2 \frac{(n'_2/n_2)(1-n'_2/n_2)}{n_2 - 1}\right]$$

On making substitutions, it gives

$$v(\hat{\pi}_g) = \frac{1}{(.8-.2)^2}\left[(1-.2)^2 \frac{(55/160)(1-55/160)}{160-1} + (1-.8)^2 \frac{(68/140)(1-68/140)}{140-1}\right]$$

$$= .002722$$

The required confidence interval will, therefore, be

$$\hat{\pi}_g \pm 2\sqrt{v(\hat{\pi}_g)}$$

$$= .2964 \pm .1043$$

$$= .1921, .4007$$

The confidence limits obtained above, indicate that the candidate A is most likely to secure 19.21% to 40.07% of the total votes. The survey thus indicates a probable win for candidate B. ■

Moors (1971) and Tracy and Mangat (1995a) have advocated the estimation of proportion of the nonsensitive character by asking direct question to the respondents in one of the two samples, and collecting information on both the characters, through randomization device, from the other sample. Folsom *et al.* (1973) and Tracy and Mangat (1995b) used two unrelated questions instead of one.

13.6.2 *Case II - π_y Known*

Consider the survey where the units of the population are the employees of a university and the unrelated nonsensitive question is: "Were you born in the month of October ?". The proportion π_y of the employees born in the month of October can be had from their records available in the university. When π_y is known, the U-model gives more precise estimates. In such a case, only one sample of the respondents is required. The results for this situation are quite simple and are given below :

Unbiased estimator of population proportion π :

$$\hat{\pi}_g = \frac{(n'/n) - (1-p)\pi_y}{p} \tag{13.28}$$

Variance of estimator $\hat{\pi}_g$:

$$V(\hat{\pi}_g) = \frac{\theta(1-\theta)}{np^2} \tag{13.29}$$

where $\theta = p\pi + (1\text{-}p)\,\pi_y$.

Unbiased estimator of variance $V(\hat{\pi}_g)$:

$$v(\hat{\pi}_g) = \frac{(n'/n)(1-n'/n)}{(n-1)\,p^2} \tag{13.30}$$

The optimal p is chosen as close to 1 as it is possible and practicable. However, the choice of π_y is made as in π_y- unknown case. Some other modifications of the U-model for π_y known case, are due to Mangat *et al.* (1992) and Singh *et al.* (1993).

Example 13.7

Abernathy *et al.* (1970) conducted a survey in North Carolina to estimate the proportion of women having had an abortion during the past year, among white women of age 18-44 years. They drew a WR random sample of 782 white women. To each women, included in the sample, was provided a randomization device carrying following two statements :

1. "I was pregnant at some time during the past 12 months, and had an abortion which ended the pregnancy"
2. "I was born in the month of April".

The randomization device used consisted of a small, transparent, sealed plastic box. Inside the box, there were 35 red and 15 blue balls. The respondent was asked to shake the box of balls thoroughly, and to tip the box allowing one of the freely moving balls to appear in a window which was clearly visible to the respondent. If a red ball appeared, she was required to answer question (1), and if a blue ball appeared she answered question (2). The respondent's reply was simply "yes" or "no" without specifying to which question the answer referred. On completion of survey, it was found

that 27 "yes" answers had been reported. Estimate the proportion in question, and work out the confidence interval for it, taking the proportion of women born in April as .0826. This figure was obtained from a distribution of births occurring to North Carolina residents during 1924-1950.

Solution

From the statement of the example we have n = 782, p = 35/50 = .7, n' = 27, and π_y = .0826, so that, n'/n = 27/ 782 = .03453. The required estimate of proportion of women who had had an abortion during the past year, is provided by (13.28). Thus,

$$\hat{\pi}_g = \frac{(n'/n) - (1-p)\pi_y}{p}$$

$$= \frac{.03453 - (1-.7)\ (.0826)}{.7}$$

$$= .01393$$

We now work out the estimate of variance which is given by (13.30). This expression is

$$v(\hat{\pi}_g) = \frac{(n'/n)\ (1-n'/n)}{(n-1)\ p^2}$$

On substituting different values, one gets

$$v(\hat{\pi}_g) = \frac{(.03453)\ (1-.03453)}{(782-1)\ (.7)^2} = .0000871$$

Then, we compute confidence limits following (2.8). These limits will, therefore, be given by

$$\hat{\pi}_g \pm 2\sqrt{v(\hat{\pi}_g)}$$

$$= .01393 \pm 2\sqrt{.0000871}$$

$$= .01393 \pm .01867$$

$$= -.00474,\ .03260$$

Leaving inadmissible values attained by lower limit, the confidence interval for the proportion of women in the age group of 18-44 years who had had an abortion during the past year, will thus be [0, .03260]. ■

The foregoing RR methods are used for estimation of proportion for stigmatized attributes which are essentially qualitative in nature. In practice, however, one may also have to deal with quantitative sensitive characters. For instance, one may be interested in estimating the average number of induced abortions among the females in a region, or the amount of tax evaded by a particular section of society. The unrelated question

model, described earlier, was suitably modified by Greenberg *et al.* (1971) to deal with *quantitative sensitive variables.*

13.7 ESTIMATION OF MEAN FOR QUANTITATIVE CHARACTERS

Consider a sensitive variable x which is supposed to be continuous with true density g(.) , and y is an unrelated nonsensitive continuous variable with true density h(.) which is roughly similar to that of x. For example, the variable x may be monthly expenditure on hard liquor, whereas y is the monthly expenditure on vegetables (or milk) in the household. The problem is to estimate μ_x, the population mean of x. First, we consider the situation where the population mean of nonsensitive variable y is unknown.

13.7.1 *Case I - μ_y Not Known*

Analogous to U-model of Greenberg *et al.* (1969), two simple random samples of n_1 and n_2 respondents are drawn, so that, $n_1 + n_2 = n$ is the required sample size. In this case also, two randomization devices R_1 and R_2 are needed. Each device consists of two questions- one regarding character x, and the other regarding character y. These two questions could be :

1. "What is the amount of money spent on hard liquor per month in the household?"
2. "How much money do you spend on vegetables in the household during a month ?"

These questions are represented with probabilities p_i and $(1-p_i)$ respectively in the device R_i, i = 1, 2, such that $p_1 \neq p_2$. The device R_1 is used for the respondents in the first sample and the device R_2 for the respondents in the second sample. Each respondent is required to draw one statement randomly, unobserved by the interviewer, and report the answer concerning the variable to which the selected statement points. The respondent's reply is simply a number, without specifying to which question the answer refers.

Let the randomized responses from the first sample be denoted by z_{1j}, j = 1, 2,..., n_1, and those from the second sample by z_{2j}, j = 1, 2, ..., n_2. Define

$$\left.\begin{aligned} \bar{z}_1 &= \frac{1}{n_1}\sum_{j=1}^{n_1} z_{1j}, \qquad s_{1z}^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(z_{1j} - \bar{z}_1)^2 \\ \bar{z}_2 &= \frac{1}{n_2}\sum_{j=1}^{n_2} z_{2j}, \qquad s_{2z}^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(z_{2j} - \bar{z}_2)^2 \end{aligned}\right] \tag{13.31}$$

The population variances σ_{1z}^2 and σ_{2z}^2 for the randomized responses z_1 and z_2 are obtained as

$$\sigma_{iz}^2 = p_i \sigma_x^2 + (1-p_i)\sigma_y^2 + p_i(1-p_i)(\mu_x - \mu_y)^2, \quad i = 1, 2, \tag{13.32}$$

where σ_x^2 and σ_y^2 are the population variances of x and y respectively.

Unbiased estimator of population mean μ_x :

$$\hat{\mu}_x = \frac{(1-p_2)\,\bar{z}_1 - (1-p_1)\bar{z}_2}{p_1 - p_2}, \qquad p_1 \neq p_2 \tag{13.33}$$

Variance of estimator $\hat{\mu}_x$:

$$V(\hat{\mu}_x) = \frac{1}{(p_1 - p_2)^2}\left[(1-p_2)^2\,\frac{\sigma_{1z}^2}{n_1} + (1-p_1)^2\,\frac{\sigma_{2z}^2}{n_2}\right] \tag{13.34}$$

Estimator of variance $V(\hat{\mu}_x)$:

$$v(\hat{\mu}_x) = \frac{1}{(p_1 - p_2)^2}\left[(1-p_2)^2\,\frac{s_{1z}^2}{n_1} + (1-p_1)^2\,\frac{s_{2z}^2}{n_2}\right] \tag{13.35}$$

The optimal design of a randomized response survey, involving quantitative variables, also requires the appropriate choice of p_1 and p_2, wise selection of a nonsensitive variable y, and efficient allocation of total sample size into n_1 and n_2. The rules for choosing optimal values of these parameters are given below :

1. A good working rule is to select one of the p_i, i = 1, 2, close to zero and other close to 1, such that $p_1 + p_2 = 1$.
2. The unrelated character y should be chosen, such that μ_y is close to μ_x and σ_y^2 is small. However, too small a choice of σ_y^2 , compared to σ_x^2, might meet with suspicion and affect likelihood of co-operation. Therefore, keeping both efficiency and protection of privacy in mind, a choice of σ_y^2 as close as possible to σ_x^2 should be attempted.
3. Analogous to (13.27), the $V(\hat{\mu}_x)$ is minimum when the total fixed sample size is allocated according to the relation

$$\frac{n_1}{n_2} = \frac{(1-p_2)\,\sigma_{1z}}{(1-p_1)\,\sigma_{2z}} \tag{13.36}$$

with σ_{iz}^2, i = 1, 2, defined in (13.32).

Example 13.8

A survey was carried out for estimating the average number of induced abortions per woman in a small town. Two independent samples of sizes $n_1 = 24$ and $n_2 = 22$ women were drawn from the population of 1200 women in the child bearing age. The two randomization devices consisted of decks of cards bearing statements :

1. "How many abortions did you have during your lifetime ?"
2. "How many children do you think a woman should have ?"

The statements (1) and (2) occur with probabilities .8 and .2 in deck I, and with probabilities .2 and .8 in deck II.

The deck I was used for the respondents in the first sample, and deck II was employed for the second sample. Each woman in the two samples was asked to draw a statement at random, unobserved by the interviewer, and report the answer concerning the selected statement without revealing as to which question the answer referred. The information thus collected is shown in table 13.3.

Table 13.3 The responses obtained from the sampled women

Sample I				Sample II			
Woman	Z_1	Woman	Z_1	Woman	Z_2	Woman	Z_2
1	2	13	2	1	2	12	3
2	1	14	3	2	2	13	2
3	2	15	1	3	1	14	2
4	1	16	1	4	3	15	1
5	2	17	2	5	1	16	2
6	1	18	1	6	2	17	2
7	2	19	2	7	2	18	3
8	3	20	1	8	2	19	2
9	0	21	2	9	3	20	2
10	2	22	3	10	1	21	2
11	1	23	1	11	2	22	2
12	3	24	3				
Total	20		22		21		23

Estimate the average number of abortions a woman had, and construct approximately 95% level confidence interval for it.

Solution

We have $n_1 = 24$, $n_2 = 22$, so that, $n = 24 + 22 = 46$. Also, $p_1 = .8$ and $p_2 = .2$. Let us first compute

$$\bar{z}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} z_{1j} = \frac{20+22}{24} = 1.75$$

$$\bar{z}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} z_{2j} = \frac{21+23}{22} = 2.00$$

$$s_{1z}^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1} (z_{1j} - \bar{z}_1)^2$$

$$= \frac{1}{24-1}[(2-1.75)^2 + (1-1.75)^2 + \ldots + (3-1.75)^2]$$

$$= .7174$$

$$s_{2z}^2 = \frac{1}{22-1}[(2-2)^2+(2-2)^2+\ldots+(2-2)^2]$$
$$= .3810$$

The estimate of average number of induced abortions is computed by using (13.33). That gives

$$\hat{\mu}_x = \frac{(1-p_2)\bar{z}_1-(1-p_1)\bar{z}_2}{p_1-p_2}$$
$$= \frac{(1-.2)(1.75)-(1-.8)(2.00)}{.8-.2}$$
$$= 1.667$$

We now compute the estimate of variance $V(\hat{\mu}_x)$ using (13.35). Thus,

$$v(\hat{\mu}_x) = \frac{1}{(p_1-p_2)^2}\left[(1-p_2)^2\frac{s_{1z}^2}{n_1}+(1-p_1)^2\frac{s_{2z}^2}{n_2}\right]$$

On substituting different values, it becomes

$$v(\hat{\mu}_x) = \frac{1}{(.8-.2)^2}\left[(1-.2)^2\left(\frac{.7174}{24}\right)+(1-.8)^2\left(\frac{.3810}{22}\right)\right]$$
$$= .05506$$

The confidence limits within which the population average is likely to fall are worked out as

$$\hat{\mu}_x \pm 2\sqrt{v(\hat{\mu}_x)}$$
$$= 1.667 \pm 2\sqrt{.05506}$$
$$= 1.667 \pm .469$$
$$= 1.198,\ 2.136$$

Thus, the average number of induced abortions per woman, in the population of 1200 women, is likely to fall in the interval [1.198, 2.136]. ■

13.7.2 *Case II - μ_y Known*

In case of binomial responses, we have seen that the estimates from the survey could be made more efficient when the value of π_y for the neutral question was known in advance. To apply this principle to quantitative responses, one should choose a nonsensitive variable for which the population mean is known in advance. For example, the question related to nonsensitive variable might, in some cases, ask for the number of persons living in a household where average household size is known from some kind of census or previous studies.

When μ_y is known, analogous to Greenberg *et al.* (1969) model, here also only one sample is required. As in case of U-model, there is a substantial reduction in variance $V(\hat{\mu}_x)$ and the results get quite simplified.

Unbiased estimator of mean μ_x :

$$\hat{\mu}_x = \frac{\bar{z} - (1-p)\mu_y}{p} \tag{13.37}$$

Variance of estimator $\hat{\mu}_x$:

$$V(\hat{\mu}_x) = \frac{\sigma_z^2}{np^2} \tag{13.38}$$

Estimator of variance $V(\hat{\mu}_x)$:

$$v(\hat{\mu}_x) = \frac{s_z^2}{np^2} \tag{13.39}$$

Terms s_z^2 and σ_z^2 above are defined in (13.31) and (13.32).

When estimating mean for a sensitive quantitative variable x through the technique discussed in this section, it is desirable that the density functions g(.) and h(.) for the sensitive variable x and the nonsensitive variable y respectively, are similar. This helps in protecting the confidentiality of the respondents, and thus results in enhanced co-operation from them. In some cases, one may be able to find a suitable nonsensitive variable y with a density function similar to that of x, whereas in other cases no such variable can possibly be found. In such situations, one may make use of the fact that any density function can be approximated by a frequency table with k class intervals, where class frequencies are proportional to the corresponding areas obtained from the density function under consideration. One can then prepare a set of cards where on cards numbering equal to the frequency for the i-th class interval is written a number equal to the mid-point of that class interval for i = 1, 2, ..., k. These numbers on the cards will be treated as the values of the nonsensitive variable y. Thus, the total number of cards in the set will be equal to the total frequency in the frequency table. This set of cards is then mixed with another set of cards bearing the statement "What is the value of x for you ?". The two sets of cards together constitute a deck of cards which can be used as the randomization device. Total number of cards in the set bearing the statement on the sensitive variable x, is determined in such a way that their proportion in the deck is equal to p. Each sample respondent, on being instructed by the investigator, would then randomly draw a card, unobserved by the investigator, from this deck. Depending on the statement on the card drawn, the respondent will report either the value of the sensitive variable x for himself or the number written on the card.

The mean value μ_y and the variance σ_y^2 for the numbers written on the cards in the set prepared from the frequency table (which are assumed to be the values taken by the nonsensitive variable y) can be easily obtained by the investigator.

The data for μ_y known case, when the unrelated characteristic with density function similar to that of the study variable is available, can be analyzed in a straightforward manner. Below we consider an example, where searching of such an unrelated

characteristic is difficult. Instead, a density function roughly close to that of the study variable is formulated by the investigator. This density function is then used in constructing the randomization device.

Example 13.9

The Income Tax Department is interested in estimating average annual income of advocates working in a court of law. These advocates number 400. The investigator is able to formulate rough distribution of the study variable which is given in table 13.4 below.

Table 13.4 Rough distribution of the income (in '000 rupees) of advocates

Total income	Mid-points	Advocates	Relative frequency	Cards prepared
48-52	50	10	.025	5
53-57	55	24	.060	12
58-62	60	60	.150	30
63-67	65	108	.270	54
68-72	70	88	.220	44
73-77	75	64	.160	32
78-82	80	24	.060	12
83-87	85	10	.025	5
88-92	90	6	.015	3
93-97	95	4	.010	2
98-102	100	2	.005	1
Total		400		200

A set of 200 cards was then prepared by writing on them, the income equal to the mid points of the class intervals in the above frequency table. Number of cards bearing a particular number were proportional to the relative frequency in table 13.4. For instance, 5 cards carried the statement "Give your response as rupees 50 thousand", whereas 12 cards carried the statement "Give your response as rupees 55 thousand", and so on. Besides these 200 cards, 600 cards bore the statement "What is your actual income ?". Each respondent in a WR simple random sample of 28 advocates, selected a card randomly from the randomization device consisting of 800 cards, and reported a number in accordance with the statement selected (near multiple of 5 thousand if actual income is to be reported). The responses, so obtained, are given in table 13.5.

Table 13.5 The responses obtained from the selected advocates

Advocate	Response	Advocate	Response	Advocate	Response	Advocate	Response
1	70	8	60	15	80	22	65
2	80	9	55	16	75	23	60
3	50	10	80	17	65	24	65
4	55	11	60	18	100	25	90
5	100	12	95	19	90	26	100
6	75	13	60	20	50	27	65
7	50	14	70	21	85	28	70

Estimate the average annual income of an advocate, and place confidence limits on it.

Solution

The statement of the problem provides n = 28 and p = 600/800 = .75. The average annual income of an advocate from the rough distribution is computed by using columns (2) and (3) of table 13.4. Thus,

$$\mu_y = \frac{1}{400}[(10)(50) + (24)(55) + \ldots + (2)(100)]$$

$$= 68.225$$

Before proceeding to obtain the required estimate, we work out the randomized response mean $\bar{z}$ for the sample units. Thus,

$$\bar{z} = \frac{1}{n}\sum_{j=1}^{n} z_j$$

$$= \frac{1}{28}(70 + 80 + \ldots + 70)$$

$$= 72.143$$

Estimate of average annual income of an advocate is obtained by using (13.37). The expression for this estimator is

$$\hat{\mu}_x = \frac{\bar{z} - (1-p)\mu_y}{p}$$

Substituting the values of $\bar{z}$, μ_y, and p, we get

$$\hat{\mu}_x = \frac{72.143 - (1-.75)\,68.225}{.75}$$

$$= 73.449$$

In order to obtain the estimate of variance, we calculate

$$s_z^2 = \frac{1}{n-1} \sum_{j=1}^{n} (z_j - \bar{z})^2$$

$$= \frac{1}{n-1} \left(\sum_{j=1}^{n} z_j^2 - n\bar{z}^2 \right)$$

$$= \frac{1}{28-1} [(70)^2 + (80)^2 + ... + (70)^2 - (28)(72.143)^2]$$

$$= \frac{1}{27} [152450 - (28)(72.143)^2]$$

$$= 248.920$$

From (13.39), the estimator of variance is given by

$$v(\hat{\mu}_x) = \frac{s_z^2}{np^2}$$

On substituting different values, one gets

$$v(\hat{\mu}_x) = \frac{248.920}{(28)(.75)^2} = 15.8044$$

The confidence limits follow from (2.8). These are obtained as

$$\hat{\mu}_x \pm 2 \sqrt{v(\hat{\mu}_x)}$$

$$= 73.449 \pm 2 \sqrt{15.8044}$$

$$= 73.449 \pm 7.951$$

$$= 65.498, 81.400$$

It indicates that if all the 400 advocates were interviewed and had they reported their annual income (in near multiple of 5 thousand) correctly, the average annual income of an advocate, almost surely, would have taken a value in the range of 65.498 to 81.400 thousand rupees. ■

For the sake of illustration, in example 13.9 we have kept the width of class interval as 5. This could be further reduced in an actual survey so as to obtain better co-operation from the respondents, and hence elicit more accurate information. Another point to be kept in mind, while building rough distribution similar to that of the study variable, is that the end points of the frequency distribution are so chosen that possibly no value pertaining to the study variable for the survey falls outside these end points.

Remark 13.1 Several other models for estimation of mean for sensitive quantitative characters are also available. Among them, the popular ones are due to Liu *et al.* (1975) and Eichhorn and Hayre (1983).

LET US DO

13.1 What is meant by nonresponse error in sample surveys ? Which are the two methods commonly used to control its effects ?

13.2 Describe briefly the Hansen and Hurwitz technique used to reduce the nonresponse bias in mail surveys.

13.3 The interest of a car manufacturing company is to estimate the average distance run by a newly purchased car before the first free service is availed. The company's recommendation is for 500 km. The frame consists of buyers who have purchased cars during the period of past 6 months to 1 year. A simple random WOR sample of 40 buyers from a population of 780 was drawn, and the questionnaire was mailed to the sampled respondents. Of these, 24 buyers responded, and from the remaining 16 nonrespondents 10 were selected for personal interview. The information collected from the buyers regarding the distance covered by the car before they availed first free service, is given below :

Through mail				Through interview	
Buyer	Distance (in km)	Buyer	Distance (in km)	Buyer	Distance (in km)
1	609	13	463	1	636
2	836	14	768	2	885
3	450	15	740	3	490
4	490	16	621	4	510
5	670	17	550	5	712
6	860	18	462	6	806
7	1007	19	880	7	532
8	630	20	1120	8	470
9	785	21	703	9	880
10	647	22	609	10	702
11	580	23	718		
12	520	24	780		

Estimate the average distance covered by a car before first free service was availed, and place confidence limits on it.

13.4 Explain Politz and Simmons procedure for reducing the nonresponse bias without resorting to call-backs.

13.5 A private company has engaged 540 workers. It is suspected that the company is not paying the workers their due wages. The investigator working on the case managed to have a complete list of all the workers along with their residential addresses. He selected a without replacement simple random sample of 30 workers. The residences of the workers included in the sample, were visited by the investigator during late evening hours. In case the respondent was found at home, he/she was asked to report the hourly wages (in rupees) paid to him/her. The respondent was also asked the number of times he/she was at home (t), at the time of interview, during the preceding five days (excluding Sunday). The information thus collected is given below. The mark "—" indicates that the respondent was not at home at the time and date of interview, and hence no information could be collected.

Worker	Wages (in Rs)	t	Worker	Wages (in Rs)	t	Worker	Wages (in Rs)	t
1	3.50	3	11	2.75	4	21	2.25	3
2	3.24	1	12	2.50	3	22	3.00	2
3	2.00	5	13	3.00	5	23	3.00	4
4	2.00	0	14	2.00	5	24	3.25	1
5	2.25	2	15	2.75	4	25	2.00	0
6	3.00	3	16	—	—	26	3.00	2
7	2.00	5	17	3.25	5	27	2.75	4
8	3.00	4	18	2.00	4	28	2.25	5
9	3.00	3	19	—	—	29	3.00	4
10	2.00	3	20	2.50	3	30	2.00	4

Estimate average hourly wages paid to the workers, and also work out confidence interval for it.

13.6 "Randomized response technique reduces/eliminates the evasive answer bias in surveys dealing with sensitive issues". Comment on the statement.

13.7 Describe briefly the Warner's randomized response technique for estimating the prevalence of sensitive attributes.

13.8 Warner's RR technique was employed in a survey to study the prevalence of smoking among undergraduate male students in a university. A student was defined as smoker if he had consumed at least two packets of cigarettes (40 cigarettes) over the one month period preceding the interview. A WR simple random sample of 250 students was drawn from a population of 2800 students. Each interviewee, included in the sample, was provided with a randomization

device consisting of a deck of cards. This randomization device carried two mutually exclusive statements :

1. "Have you consumed at least two packets of cigarettes during the last one month period?"
2. "Did you consume less than two packets of cigarettes, or did not smoke at all during the last one month period?"

with probabilities .75 and .25 respectively. Each interviewee was required to draw a card at random after reshuffling the deck, unobserved by the interviewer, and report "yes" if the statement on the selected card points to his actual status, and "no" otherwise. There were altogether 90 "yes" and 160 "no" answers. Estimate the proportion of smokers, and set confidence limits for it.

13.9 Suppose that for the situation considered in exercise 13.8, Mangat and Singh's two-stage procedure was used. Taking $n = 250$, $p = .75$, and $T = .3$, the number of "yes" answers recorded were 84. Estimate the proportion of smokers in the population, and place confidence limits on it.

13.10 "The confidence of the respondents in the anonymity provided by Warner's pioneer RR model might be further enhanced if one of the two questions confronted by the respondents referred to a nonsensitive attribute unrelated to the sensitive attribute under study". Explain the sampling strategy based on this principle.

13.11 There are 1560 workers in a factory. The management suspects that some workers might be involved in gambling. In order to estimate the proportion of such workers, two independent samples of 70 and 65 workers were selected through SRS with replacement. Each respondent in the first sample was provided with a randomization device carrying two statements:

1. "Are you a habitual gambler ?"
2. "Are you familiar with the rules of cricket ?"

with probabilities .85 and .15. The respondents in the second sample were provided with a similar randomization device representing the statements (1) and (2) with probabilities .15 and .85 respectively. Each respondent in the first and second samples was asked to choose randomly a statement from the RR device provided to him, and to say "yes" or "no" depending on whether or not the selected statement points to his actual status. Altogether 13 "yes" answers were reported by the interviewees in the first sample, whereas 9 "yes" answers were reported from the second sample. Estimate the proportion of habitual gamblers among the workers, and also work out confidence limits for it.

13.12 The income tax department suspects that some of the teachers of a university have income from alternative sources, viz., business, tuition work, and other type of part time employment. This income is, however, not reported by teachers when the income tax is deducted at source from their salary. To estimate the proportion of such teachers, a WR simple random sample of 160 teachers was drawn from

a population of 1400. Each respondent included in the sample was provided with a randomization device. The device consisted of 200 cards of which 160 bore the statement "Do you have income from sources other than the university salary ?", and the remaining 40 cards carried the statement "Were you born in the month of October ?" The proportion of teachers born in October is known to be .0917. On completion of survey, it was found that in all there were 32 "yes" and 128 "no" responses. Estimate the proportion in question, and construct confidence interval for it.

13.13 Discuss application of RR technique in obtaining data on a quantitative sensitive variable.

13.14 A survey was undertaken to estimate the amount of money spent on alcoholic drinks by the students of a university. With this objective in view, two independent WR simple random samples of sizes 25 and 30 students were selected from a total population of 1060 students. For interviewing the respondents, two sets of randomization device were used – set 1 for the respondents in the first sample and set 2 for the second sample. Set 1 consisted of cards carrying following two questions :

1. "How much have you spent on alcoholic drinks during the last 3 months ?"
2. "How much have you spent on purchasing clothes for yourself during the last 2 months ?"

in proportions .8 and .2 respectively. The cards in set 2 carried questions (1) and (2) in proportions .2 and .8 respectively. Each respondent was required to choose randomly one of these questions and report answer with respect to the question chosen. The responses obtained, in terms of rupees, are given below :

Sample 1 :	280	170	50	0	100	40	120	80	0	0	90	190	50
	0	115	150	60	80	30	0	0	100	0	110	70	
Sample 2 :	0	110	85	220	170	60	0	150	0	170	220	0	0
	100	65	90	190	240	0	170	0	50	100	0	40	100
	60	230	0	145									

Estimate the average expenditure incurred by the students on alcoholic drinks during the last 3 months, and place confidence limits on it.

13.15 It is desired to estimate the average frequency of traffic rule violations by the university employees. To accomplish the objective, 40 employees from the total population of 950, were selected through WR simple random sampling. Each employee in the sample was provided with a randomization device carrying two statements :

1. "How many times have you been issued tickets for violating traffic rules during last five years ?"
2. "How many visits did you make to university hospital for treatment during the last two months ?"

The statements (1) and (2) were represented with probabilities .7 and .3 respectively. The average number of visits of a university employee to the hospital were worked out from the hospital records, where the file number of each employee is entered before one is treated. It came out to be 2.3. Each respondent selected a statement randomly from the device provided and answered accordingly. The responses obtained are listed below :

Responses : 2 1 0 0 3 1 2 0 4 3 0 1 4 2 5 1 3
5 1 0 2 3 1 4 5 1 0 3 2 3 1 4 0 0
5 0 1 1 0 0

Estimate average number of traffic tickets issued to a university employee, and place approximately 95% confidence level limits on it.

Appendixes

Appendix A Standard normal probability distribution

	Areas under standard normal curve (0 to Z)									
Z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4990	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4966	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

Source : Rao, C.R., Mitra, S.K., Matthai, A., and Ramamurthy, K.G. (1974). *Formulae and Tables for Statistical Work.* Statistical Publishing Society, Indian Statistical Institute, Calcutta.

Appendix B Random numbers

Column numbers									
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
3436	6833	5809	9169	5081	5655	6567	8793	6830	1332
6133	4454	2675	3558	7624	5736	2184	4557	0496	8547
9853	3890	5535	3045	9830	5455	8218	9090	7266	4784
5807	5692	6971	6162	6751	5001	5533	2386	0004	2855
6291	0924	1298	7386	5856	2167	8299	9314	0333	8803
4725	9516	8555	0379	7746	9647	2010	0979	7115	6653
7697	6486	3720	6191	3552	1081	6141	7613	5455	3731
3497	2271	9641	0304	4425	6776	1205	2953	5669	1056
8940	4765	1641	0606	4970	7582	7991	6480	2946	5190
1122	6364	5264	1267	4027	4749	0338	8406	1213	5355
4333	0625	3947	1373	6372	9036	7046	4325	3491	8989
7685	1550	0853	4276	1572	9348	6893	2113	8285	9195
0592	8341	4430	0496	9613	2643	6442	0870	5449	8560
3506	0774	0447	7461	4459	0866	1698	0184	4975	5447
8368	2507	3565	4243	6667	8324	3063	8809	4248	1190
2630	1112	6680	4863	6813	4149	8325	2271	1963	9569
3883	3897	1848	8150	8184	1133	6088	3641	6785	0658
1123	3943	5248	0635	9265	4052	1509	1280	0953	9107
1167	9827	4101	4496	1254	6814	2479	5924	5071	1244
7831	0877	3806	9734	3801	1651	7169	3974	1725	9709
2487	9756	9886	6776	9426	0820	3741	5427	5293	3223
1245	3875	9816	8400	2938	2530	0158	5267	4639	5428
5309	4806	3176	8397	5758	2503	1567	5740	2577	8899
7109	0702	4179	0438	5234	9480	9777	2858	4391	0979
8716	7177	3386	7643	6555	8665	0768	4409	3647	9286
9499	5280	5150	2724	6482	6362	1566	2469	9704	8165
3125	4552	6044	0222	7520	1521	8205	0599	5167	1654
3788	6257	0632	0693	2263	5290	0511	0229	5951	6808
2242	2143	8724	1212	9485	3985	7280	0130	7791	6272
0900	4364	6429	8573	9904	2269	6405	9459	3088	6903
7909	4528	8772	1876	2113	4781	8678	4873	2061	1835
0379	2073	2680	8258	6275	7149	6858	4578	5932	9582
0780	6661	0277	0998	0432	8941	8946	9784	6693	2491
8478	8093	6990	2417	0290	5771	1304	3306	8825	5937
2519	7869	9035	4282	0307	7516	2340	1190	8440	6551
2472	0823	6188	3303	0490	9486	2896	0821	5999	3697
8418	5411	9245	0857	3059	6689	6523	8386	6674	7081
8293	5709	4120	5530	8864	0511	5593	1633	4788	1001
9260	1416	2171	0525	6016	9430	2828	6877	2570	4049
6568	1568	4160	0429	3488	3741	3311	3733	7882	6985
6694	5994	7517	1339	6812	4139	6938	8098	6140	2013
2273	6882	2673	6903	4044	3064	6738	7554	7734	7899
6364	5762	0322	2592	3452	9002	0264	6009	1311	5873
6696	1759	0563	8104	5055	4078	2516	1631	5859	1331
3431	2522	2206	3938	7860	1886	1229	7734	3283	8487
4842	3765	3484	2337	0587	9885	8568	3162	3028	7091
8295	9315	5892	6981	4141	1606	1411	3196	9428	3300
4925	4677	8547	5258	7274	2471	4559	6581	8232	7405
5439	0994	3794	8444	1043	4629	5975	3340	3793	6060
2031	0283	3320	1595	7953	2695	0399	9793	6114	2091

Sources : Rao, C.R., Mitra, S.K., Matthai, A. and Ramamurthy, K.G. (1974). *Formulae and Tables for Statistical Work*. Statistical Publishing Society. Indian Statistical Institute Calcutta.

Appendix B continued...

Column numbers									
(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
9787	3792	5241	0556	7070	0786	7431	7157	8539	4118
4479	1397	8435	3542	8435	6169	7996	3314	1299	1935
0191	2800	1056	2735	4816	1979	0042	5824	6636	2332
8710	6903	1347	9332	6962	6786	9875	7565	8683	6490
4656	5960	0812	5144	5355	3335	4784	7573	3841	4255
9974	9230	8049	4971	7555	3935	9405	8545	4329	5358
8493	7128	3654	8976	1901	5496	3453	7539	3255	6742
6135	6954	3436	3841	9009	3768	9256	3631	9066	7153
1217	2748	3864	4752	7407	9975	6372	3308	0000	4734
2623	1282	4389	8889	0764	2328	2140	8843	4986	4413
1144	5336	4426	9003	6956	9406	8464	8827	3143	4754
5854	9981	9079	2908	4755	4620	6455	6793	7539	4031
0615	8188	2812	0270	5733	5339	1175	2919	7343	0477
3624	0853	3128	7952	2678	3011	7710	9734	6386	8400
1185	6832	4918	9236	3026	5795	0352	7533	4435	0306
7391	3210	9540	4085	9234	4892	3962	3883	4538	8286
7195	1986	6146	0946	5421	8430	2128	7602	5609	7064
6137	7286	5283	0609	0941	4935	2521	7937	2153	2629
7401	8099	7482	2210	3662	8253	7507	7809	0094	4401
0192	9452	7189	9552	7498	0105	8295	9762	7434	3518
3621	3037	2274	3803	0946	9874	4911	6797	1227	8494
2661	0047	6628	6199	2526	5631	8334	7668	3994	7439
8072	5085	3576	4939	0352	7386	7690	7108	6668	8246
0839	5224	9768	3839	8495	1668	6957	7031	2032	1468
2354	9266	8034	3813	3648	7825	6156	3605	7796	1645
9050	6800	0490	3261	7748	3609	1050	0591	3799	2827
7174	7703	1540	8001	6230	0387	9553	7447	0240	2511
3465	7017	2278	0357	5800	1048	8382	8800	7608	4325
8805	1265	5202	6872	3282	5331	5398	1426	2805	2110
0250	4100	5263	8506	9848	2451	2031	2026	8661	4163
6088	8366	7751	1577	9534	2458	1886	1522	4161	8726
8833	3449	3499	4223	2854	6855	4042	1294	1728	5494
4675	2535	1915	9783	9754	2790	6856	0352	9628	8342
8990	4993	2922	8842	9904	8442	0105	3308	3320	6361
1790	8590	5792	0983	3494	0945	4966	2194	9823	2599
9276	3967	2486	6242	3276	1884	1847	8922	7356	1528
2965	7991	3777	9303	0536	1517	0570	7212	7593	0566
6620	4234	8407	6890	6904	8599	5876	2608	7320	6117
4706	8319	6252	3177	9108	3069	0910	8241	9842	0895
8395	3882	0259	2092	4885	3434	0879	0000	0790	0735
3991	3406	0151	2594	9137	9924	2393	7699	6116	9655
9644	6763	3512	0139	4119	2722	3219	0070	3830	7997
3658	7813	0207	0357	8225	4497	2435	5121	4776	3611
5728	1882	9120	7893	3503	8579	9070	1952	8390	5517
6221	6366	8192	8429	4387	5484	7553	4053	9458	2292
7635	5248	1750	0868	0173	4989	2300	3916	6732	8284
4368	3113	5887	8439	0026	1902	4114	3127	5140	6684
8635	9723	2550	8216	7531	7732	3963	4014	2099	3030
3304	3254	3936	9361	9771	8255	4592	8808	3803	4010
9336	5666	1349	1932	7326	2151	1573	3045	8746	8059

Appendix B continued...

Column numbers									
(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
8094	7747	6006	2536	8856	8171	0291	2603	4675	0779
3745	6766	7221	9560	7036	4520	4584	5714	8122	5029
0835	3641	1638	3464	1767	7664	6247	8362	1257	5265
1601	1143	7272	3988	8356	9477	5870	6425	1725	9792
7797	4121	9603	5723	4630	5549	0593	5761	7200	7227
7620	8310	9500	7116	6259	7619	3749	9121	2185	9335
2096	5270	0793	3950	2722	0925	5792	1040	5806	9636
6803	7016	1055	6396	7754	3591	2613	5325	7485	2406
3566	9310	2604	8607	4765	2237	1222	3947	1228	2708
6428	0086	6245	3247	5707	7847	6127	0857	8229	5609
8633	2617	9176	9602	4807	7269	6131	8780	3417	7278
6632	8056	1091	9158	7303	4084	9096	4047	6775	0876
2612	7936	1453	4812	1742	7128	3636	6561	7522	0359
9436	1681	0851	3488	8815	5301	5403	5456	0501	4511
0418	2487	5583	9032	6507	8554	0346	6251	3577	4146
6853	3757	0171	5943	1145	3434	0188	5665	7779	7179
8347	7044	4640	6832	2445	4872	7870	2335	2874	9393
5182	6263	1224	9863	6751	0084	8827	9479	8342	0053
9215	3992	4874	8082	5959	2861	4574	5813	5903	7161
5588	3456	9602	5260	6578	8618	0340	3381	7579	6359
3996	0415	7015	9210	0974	0319	2699	8036	1090	3805
7346	9400	3292	8165	3206	7035	5227	7340	8515	4225
8621	4185	6727	2770	1227	3696	6496	4889	2697	3316
9399	5575	1562	5821	9824	4909	0348	8735	3604	9959
4334	0347	4893	2025	5590	8126	8571	2532	9355	7563
8091	0536	6522	5409	1463	0138	0384	6711	2384	0072
9627	3311	2010	2525	3142	9700	2196	4076	3710	3372
0086	3501	4916	2511	1274	1775	8324	9646	0611	1048
3753	0174	7934	3483	9210	9163	4714	7888	3577	6596
2740	3239	3054	9991	3778	3195	1040	2022	3193	9196
3919	6871	5685	8147	7310	2080	4196	3375	5700	7967
4577	7897	2757	5992	7398	7687	8415	1595	9636	4605
0215	7254	5378	3861	3448	9494	5221	1325	7317	1022
5807	7948	1774	6836	1786	2392	2820	8533	0629	3771
1910	9653	1214	3921	5298	8334	2352	7113	2291	9312
3990	1310	9338	2601	5571	1424	7850	4531	0133	5519
5967	8941	7987	3335	7579	9735	3042	8409	7053	5364
5872	1143	9183	6911	2247	1559	4888	7198	9249	1395
7240	1827	3281	0705	4479	5598	9985	8170	3367	6928
2268	4227	5844	0700	6907	9668	6670	0097	0686	6311
8515	1611	1327	6671	2765	0081	0554	3716	9334	3027
4324	8348	8870	4802	9655	2852	3858	3225	5022	3602
9053	8503	8222	6850	6100	5973	1522	2690	1396	0632
5133	7618	3211	0898	5343	9981	8936	0819	9112	2548
6235	9463	0097	1332	6038	3822	1119	7143	1708	5668
6048	1376	1589	4274	2920	3521	7661	9435	9257	9276
6341	0636	3355	7245	4160	1672	2295	4730	0984	6813
2143	0207	9733	8136	9118	0143	0949	1733	7986	5670
7336	3277	2135	3300	5287	0134	7104	9359	5069	3893
2728	6464	4721	8192	5485	7935	4996	3475	9523	5514

Appendix B continued...

Column numbers									
(31)	(32)	(33)	(34)	(35)	(36)	(37)	(38)	(39)	(40)
6415	5554	3592	8008	9408	2092	9842	3197	1404	1505
4668	3479	4073	6941	8286	3374	3696	7856	8980	0359
7592	3903	7895	1113	7646	9201	9081	2630	1617	1188
2012	1096	2958	4788	4882	1855	8190	9726	6716	1384
7884	8004	7831	8264	0028	8118	5011	5704	9394	7669
5510	8160	6173	5655	4415	0147	1091	4426	2843	5578
4440	0095	4067	9078	6205	7488	1851	3537	7191	0856
8436	4936	3013	6818	1577	0249	5107	5304	3872	4157
3740	3172	2775	5781	0318	8932	9220	3784	0501	8375
1174	3869	9985	4443	1127	7390	1463	8524	2272	4275
8494	5214	9020	4568	3508	1257	9685	6310	9763	1887
8792	6689	3521	4407	2017	8527	2230	1851	4023	2258
0865	4556	4015	0082	1239	7058	1189	3174	0220	1167
7141	0799	4764	5283	4291	4822	3735	1393	2477	6782
7185	3986	7047	9210	2791	7610	7264	4771	0548	5172
3672	8714	8853	9825	5869	6281	2371	1890	9480	2968
7753	9791	3436	4604	7991	5222	9280	1584	7141	0221
9332	5082	8900	4209	4117	8644	8712	7337	1689	8793
0759	2206	4220	2394	4346	8483	6968	2344	1902	0848
8493	6032	3585	2162	6301	4929	7087	2907	2690	5039
6776	2659	7323	9619	7727	6460	6745	1051	7662	7512
5135	7118	4458	1394	0526	5121	2062	0977	7338	5744
7714	3485	5412	0716	6914	8192	6483	1946	4271	0995
9777	1915	1183	3177	6568	6698	4649	3899	2691	4413
7960	4876	8841	3538	4519	0872	5860	8181	5777	0233
1714	4061	6365	7480	9312	1139	0715	0571	2575	5990
7460	0288	1075	3483	1041	5427	6457	0985	1657	8742
0275	8595	0812	9021	4808	8247	0089	7034	8719	5878
7735	0399	3931	3135	1585	7292	8362	4006	1184	9676
8661	9964	9969	2444	6095	2003	9320	2837	4397	0297
1273	7133	4874	1100	7854	4596	6787	8574	6098	5526
7784	9159	6674	3243	2531	6093	8906	8855	8614	2781
0707	0067	6433	6058	4381	0146	1186	9913	3668	6347
9594	8627	5507	2956	6166	7271	9511	5069	1022	9889
6690	2781	1790	9596	6472	8774	9058	7915	3647	3525
3476	7990	0690	0043	1357	9568	1541	3726	9223	4385
9994	1061	7951	3010	6997	4759	0473	2848	7504	6904
8308	8100	7244	4206	7766	6916	6866	4064	6714	1805
7260	8057	8779	6368	0601	1872	3160	8731	3646	2789
4755	3425	1299	7990	8366	1368	3611	8864	1341	9349
7156	7190	6054	3489	8939	9089	2637	9180	3991	7161
1469	1763	1918	2547	7708	1900	1665	1860	3078	7851
1270	4109	9428	0933	1444	7467	1771	3482	1497	6492
5485	7802	3094	7249	3901	2827	8294	1329	7170	1758
7123	0850	6297	5479	1416	1837	9305	3749	8541	5161
2187	4696	2470	7234	4809	5408	3266	6252	5987	5794
7595	1895	6183	2013	4399	5255	6714	1839	6132	2653
3021	1523	2005	2009	9631	1274	9902	4203	8312	9572
3317	8741	2688	9392	0136	9293	7815	1781	1990	4057
6711	3947	5004	2625	5105	0116	1895	6729	3159	6492

Appendix C Number of tractors, tube wells, and net irrigated area (in hectares) for 69 villages of Doraha development block of Punjab, India

Village No.	Village name	Tractors	Tube wells	Irrigated area
1	Ajnaud	20	102	281
2	Aracha	19	126	337
3	Afjullapur	5	31	77
4	Alampur	10	39	108
5	Bishanpur	12	85	191
6	Shahpur	21	130	208
7	Sirthala	44	222	698
8	Sultanpur	12	70	180
9	Sihora	20	219	458
10	Hole	10	115	288
11	Kotli Afgana	11	41	161
12	Katari	30	133	512
13	Kartarpur	15	55	123
14	Katana Sahib	3	50	100
15	Kaddon	36	249	675
16	Gobindpura	24	119	258
17	Gurditpura	17	95	143
18	Gidrhi	15	100	267
19	Ghadani Kalan	76	551	1178
20	Ghadani Khurd	35	209	523
21	Ghalouti	46	380	583
22	Ghanrash	16	160	310
23	Chankoian Kalan	7	33	129
24	Chankoian Khurd	21	142	184
25	Chapran	7	64	91
26	Jaipura	22	130	240
27	Jahangir	5	27	80
28	Jandali	21	118	429
29	Jargarhi	38	193	583
30	Jarag	59	360	888
31	Jallah	17	94	403
32	Taunsa	15	60	152
33	Turmusi	8	62	145
34	Dipnagar	14	70	345
35	Daburjee	18	116	346
36	Dugri	12	115	261
37	Diva - Mander	15	95	118
38	Doraha	14	97	467
39	Dhamot	86	595	1723
40	Nizampur	26	114	310
41	Ferozepur	2	81	110
42	Fatehpur	4	28	150
43	Buani	19	106	479
44	Barmalipur	29	191	328
45	Bivipur	13	71	197
46	Begowal	20	204	433

Source : *Directory of Villages - District Ludhiana.* Punjab Government, India (1983-84).

Appendix C continued...

Village No.	Village name	Tractors	Tube wells	Irrigated area
47	Bishanpura	9	33	393
48	Baupur	8	47	127
49	Bilaspur	38	175	409
50	Bharthala	10	65	217
51	Bhathal	15	61	177
52	Mahnpur	8	48	222
53	Mangewal	12	100	284
54	Modnipur	11	111	236
55	Maksudra	42	320	761
56	Malipur	8	103	205
57	Majree	23	115	246
58	Mulanpur	15	108	120
59	Malakpur	12	88	173
60	Rara Sahib	9	67	115
61	Raul	8	51	215
62	Rampur	39	393	1083
63	Rauni	65	370	992
64	Rano	6	90	153
65	Rajgarh	51	138	578
66	Lasara (L)	20	110	114
67	Lasara (P)	12	70	284
68	Landha	25	131	269
69	Lapran	30	92	234
Total		1465	9333	23857

Appendix D Fifty WOR simple random samples

Sample	Sample observations																		
1	29	20	20	8	30	22	11	15	21	12	12	7	14	19	9	17	38	86	23	39
2	10	8	18	17	7	29	11	36	21	20	46	8	86	10	51	12	8	35	21	29
3	15	15	21	12	5	30	12	12	15	76	59	10	8	25	2	14	7	38	86	4
4	5	9	86	21	11	8	12	12	15	10	8	15	9	21	20	17	51	38	18	12
5	8	10	15	12	12	14	14	19	29	21	17	38	8	16	20	18	10	36	16	14
6	2	18	8	9	30	38	11	12	6	22	21	15	26	20	14	19	65	15	36	16
7	12	18	5	15	38	14	9	38	19	26	20	21	16	8	15	5	25	36	4	39
8	8	4	18	15	29	2	8	5	12	12	20	20	10	86	15	8	20	23	21	7
9	8	30	12	76	44	9	23	22	59	11	20	12	7	5	39	15	15	21	14	12
10	18	24	39	42	20	20	44	6	8	8	22	14	46	2	20	8	8	59	11	10
11	44	8	76	12	8	86	11	15	7	20	12	59	23	8	38	20	8	24	12	5
12	30	18	25	21	21	35	86	11	59	12	12	17	25	7	20	8	15	15	42	4
13	9	14	65	46	6	11	5	86	7	38	30	8	10	12	35	15	26	25	20	10
14	17	4	12	12	8	26	6	42	10	8	39	21	46	44	11	14	23	10	4	20
15	8	15	12	12	38	14	12	3	35	36	30	5	46	22	15	76	26	86	17	59
16	23	3	21	51	13	86	17	12	19	5	76	14	8	19	15	20	39	25	12	9
17	30	20	14	19	15	38	17	65	29	13	8	46	30	86	35	15	19	21	2	17
18	86	12	2	38	42	20	19	18	14	20	12	14	5	25	15	26	12	12	8	21
19	2	76	7	6	4	12	12	9	29	26	10	5	22	18	3	24	21	7	19	46
20	15	18	15	42	15	4	12	14	15	11	17	6	38	20	15	86	20	25	13	5
21	12	24	21	15	19	15	15	8	29	21	8	38	36	9	12	76	20	16	20	3
22	2	9	26	39	8	76	14	36	7	5	11	4	8	11	13	3	19	46	15	18
23	22	30	51	29	15	6	25	8	2	19	21	20	5	24	16	21	20	10	86	10
24	13	8	12	15	8	17	30	19	21	38	10	65	76	12	15	11	38	29	5	7
25	12	26	36	10	19	16	76	30	30	21	15	15	21	15	44	10	8	12	51	65
26	4	20	51	21	5	22	21	17	12	46	12	13	16	11	38	7	25	12	15	3
27	8	39	30	3	76	5	25	42	21	9	11	51	17	20	76	12	11	18	46	6
28	8	20	12	16	20	30	8	25	8	12	23	7	4	11	15	38	9	65	36	12
29	86	18	14	12	65	2	21	21	12	19	15	8	20	59	76	15	21	8	36	38
30	4	19	26	8	10	15	46	29	7	5	20	76	21	17	65	39	20	12	16	51
31	36	15	14	5	15	21	4	39	8	11	8	20	38	23	16	25	44	11	5	29
32	42	8	11	20	29	51	12	12	18	65	24	1	15	13	8	16	30	8	9	38
33	21	86	8	39	15	11	4	9	59	7	13	10	23	26	15	3	76	25	12	46
34	10	5	15	18	2	12	6	8	12	14	12	15	15	44	19	11	51	9	86	76
35	9	24	7	9	4	8	36	14	19	15	35	8	15	38	13	29	20	46	76	10
36	2	16	20	7	22	8	12	12	17	20	24	36	21	23	35	86	15	13	20	10
37	76	38	15	15	5	20	46	12	12	12	17	38	29	10	15	30	22	11	19	13
38	30	23	14	65	29	4	20	8	20	17	15	76	10	14	11	22	12	10	20	44
39	12	2	12	15	14	3	7	8	15	20	10	23	10	16	6	86	21	25	14	20
40	39	19	22	23	76	20	7	15	38	8	8	17	20	86	17	21	12	7	8	15
41	12	17	23	5	10	38	17	21	26	76	30	39	12	42	10	21	19	2	20	14
42	4	25	38	65	12	44	30	19	12	7	9	13	9	39	15	15	20	18	30	35
43	26	59	8	76	15	11	25	35	17	6	30	12	14	19	10	12	21	17	2	65
44	9	21	12	14	38	16	23	10	21	10	15	19	26	51	17	5	18	20	44	8
45	13	10	38	15	20	17	20	5	26	2	10	3	12	19	86	39	10	9	14	42
46	5	18	36	46	20	15	9	15	17	10	8	24	4	15	5	21	10	38	21	9
47	12	20	12	17	29	21	8	39	19	26	20	22	15	13	12	14	38	11	30	24
48	17	2	22	16	9	12	29	8	9	12	21	15	15	8	5	15	8	12	10	15
49	4	5	1	12	11	15	8	76	12	25	3	15	12	12	30	59	13	42	39	15
50	29	35	10	8	12	59	23	5	15	6	20	19	20	9	10	8	17	5	7	15

Appendix E Explanation of certain local terms used

For political, developmental, and administrative reasons, India is divided into a number of states. Each state is further divided into different districts, the districts into tahsils, and the tahsils into development blocks. Development blocks have large number of villages in their jurisdiction. Each village belongs to one and only one development block.

Certain other local terms that appear in the book are explained below :

1. **Anganwadi** : It is a kind of nursery school/ day care center for the children in rural areas. These are run under a government welfare scheme, and are almost free of charge.
2. **Neem** : It is a tree with Azadirachta indica as its botanical name. The tree is important because of its medicinal values. This fact is known to Indian masses for thousands of years.
3. **Panchayat** : The persons belonging to a village or a group of contiguous villages elect a certain number of persons. These elected persons constitute a body which is called panchayat. It looks after the welfare of the villages.
4. **Patwari** : Patwari is the lowest revenue official who keeps records about land ownership and its usage in rural areas. He has a village or a group of adjoining villages in his charge.
5. **Ramayana** : It is a very famous and popular Hindu religious epic of India. It is believed to have been written in fifth century. It describes the life of Lord Rama in detail. It has been the subject of stage plays and films from time immemorial. Based on it, a television serial was also made. This serial was quite popular among Indian masses.
6. **Rupee** : It is the Indian currency. It is abbreviated as Re in singular form and Rs in plural form.
7. **Sarpanch** : The chairperson of the panchayat is called Sarpanch. He/she is elected directly by the villagers in the jurisdiction of panchayat.
8. **Yoga** : A system of ascetic philosophy involving religious and abstract meditation on the Supreme Spirit. It also includes certain physical body postures and exercises which are believed to prevent and cure certain diseases and disorders.
9. **Yoga Ashram** : The place or building where yoga is practiced.

References

[1] Abernathy, J.R., Greenberg, B.G., and Horvitz, D.G. (1970) Estimates of induced abortion in urban North Carolina. *Demography*, **7**, 19-29.

[2] Abul-Ela, A.L.A., Greenberg, B.G., and Horvitz, D.G. (1967) A multi-proportions randomized response model. *J. Amer. Statist. Assoc.*, **62**, 990-1008.

[3] Adams, L. (1957) *An Analysis of a Population of Snowshoe Hares.* Ph.D. Thesis. The Johns Hopkins University, Baltimore.

[4] Aoyama, H. (1954) A study of stratified random sampling. *Ann. Inst. Statist. Math.*, **6**, 1-36.

[5] Badaloni, M. (1993) Spatial sampling of sparse biological populations. *Metodi Statist. Anal. Territor.*, 30-66.

[6] Bailey, N.T.J. (1951) On estimating the size of mobile populations from recapture data. *Biometrika*, **38**, 293-306.

[7] Bansal, M.L. and Singh, Ravindra (1985) An alternative estimator for multiple characteristics in PPS sampling. *J. Statist. Planning Infer.*, **11**, 313-320.

[8] Beale, E.M.L. (1962) Some uses of computers in operational research. *Industrielle Organisation*, **31**, 51-52.

[9] Best, D.J. (1974) The variance of the inverse binomial estimator. *Biometrika,* **61**, 385-386.

[10] Bowley, A. L. (1926) Measurement of the precision attained in sampling. *Bull. Inter. Statist. Inst.*, **22**, 1-62.

[11] Brewer, K.R.W. (1963) A model of systematic sampling with unequal probabilities. *Austral. J. Statist.*, **5**, 5-13.

[12] Chakravarti, I.M., Laha, R.G., and Roy, J. (1967) *Hand Book of Methods of Applied Statistics.* Vol II. John Wiley & Sons, Inc., New York.

[13] Chapman, D.G. (1951) Some properties of the hypergeometric distribution with applications to zoological sample censuses. *University of California Publications in Statistics*, **1**, 131-160.

[14] Chapman, D.G. (1952) Inverse, multiple and sequential sample censuses. *Biometrics*, **8**, 286-306.

[15] Cochran, W.G. (1939) The use of analysis of variance in enumeration by sampling. *J. Amer. Statist. Assoc.,* **34**, 492-510.

[16] Cochran, W.G. (1977) *Sampling Techniques.* John Wiley & Sons, Inc., New York.

[17] Cox, D.R. (1952) Estimation by double sampling. *Biometrika*, **39**, 217-227.

[18] Dalenius, T. and Gurney, M. (1951) The problem of optimum stratification, II. *Skand. Akt.*, **34,** 133-148.

[19] Dalenius, T. and Hodges, J.L., Jr. (1957) The choice of stratification points. *Skand. Akt.*, **40**, 198-203.

[20] Dalenius, T. and Hodges, J. L., Jr. (1959) Minimum variance stratification. *J. Amer. Statist. Assoc.*, **54**, 88-101.

[21] Das, A.C. (1951) On two phase sampling and sampling with varying probabilities. *Bull. Inter. Statist. Inst.*, **33**, 105-112.

[22] Deming, W.E. (1953) On a probability mechanism to attain an economic balance between the resultant error of response and the bias of nonresponse. *J. Amer. Statist. Assoc.*, **48**, 743-772.

[23] Deming, W.E. (1960) *Sample Design in Business Research.* John Wiley & Sons, Inc., New York.

[24] Des Raj (1956) Some estimators in sampling with varying probabilities without replacement. *J. Amer. Statist. Assoc.*, **51**, 269-284.

[25] Des Raj (1965) On a method of using multi-auxiliary information in sample surveys. *J. Amer. Statist. Assoc.*, **60**, 270-277.

[26] Des Raj (1968) *Sampling Theory.* McGraw-Hill Book Company, New York.

[27] Des Raj and Khamis, S.H. (1958) Some remarks on sampling with replacement. *Ann. Math. Statist.*, **29**, 550-557.

[28] Dey, A. and Srivastava, A.K. (1987) A sampling procedure with inclusion probabilities proportional to size. *Survey Meth.*, **13**, 85-92.

[29] Diana, G. (1993) A class of estimators of the population mean in stratified random sampling. *Statistica,* **53**, 59-66.

[30] Durbin, J. (1954) Non-response and call-backs in surveys. *Bull. Inter. Statist. Inst.,* **34**, 72-86.

[31] Durbin, J. (1967) Design of multi-stage surveys for the estimation of sampling errors. *Appl. Statist.,* **16**, 152-164.

[32] Eichhorn, B.H. and Hayre, L.S. (1983) Scrambled randomized response methods for obtaining sensitive quantitative data. *J. Statist. Planning Infer.,* **7**, 307-316.

[33] El-Badry, M.A. (1956) A sampling procedure for mailed questionnaires. *J. Amer. Statist. Assoc.,* **51**, 209-227.

[34] Finney, D.J. (1949) On a method of estimating frequencies. *Biometrika*, **36**, 233-234.

[35] Fisher, R.A. and Yates, F. (1938) *Statistical Tables for Biological, Agricultural and Medical Research.* Table XXXIII, Oliver and Boyd, London.

[36] Fisz, M. (1963) *Probability Theory and Mathematical Statistics*. John Wiley & Sons, Inc., New York.

[37] Folsom, R.E., Greenberg, B.G., Horvitz, D.G., and Abernathy, J.R. (1973) The two alternate questions randomized response model for human surveys. *J. Amer. Statist. Assoc.*, **68**, 525-530.

[38] Franklin, L.A. (1989) Randomized response sampling from dichotomous populations with continuous randomization. *Survey Meth.*, **15**, 225-235.

[39] Ghangurde, P.D. and Rao, J.N.K. (1969) Some results on sampling over two occasions. *Sankhyā (A)*, **31**, 463-472.

[40] Ghosh, B. (1947) Double sampling with many auxiliary variates. *Calcutta Statist. Assoc. Bull.*, **1**, 91-93.

[41] Goodman, R. and Kish, L.(1950) Controlled selection - a technique in probability sampling. *J. Amer. Statist. Assoc.*, **45**, 350-372.

[42] Greenberg, B.G., Abul-Ela, A.L.A., Simmons,W.R., and Horvitz, D.G. (1969) The unrelated question randomized response model : Theoretical framework. *J. Amer. Statist. Assoc.*, **64**, 520-539.

[43] Greenberg, B.G., Kuebler, R.R., Abernathy, J.R., and Horvitz, D.G. (1971) Application of randomized response technique in obtaining quantitative data. *J. Amer. Statist. Assoc.*, **66**, 243-250.

[44] Gupta, P.C. and Adhvaryu, D. (1982) On some unbiased product type strategies. *J. Indian Soc. Agric. Statist.*, **34**, 48-54.

[45] Haldane, J.B.S. (1946) On a method of estimating frequencies. *Biometrika*, **33**, 222-225.

[46] Hansen, M.H. and Hurwitz, W.N. (1943) On the theory of sampling from finite populations. *Ann. Math. Statist.*, **14**, 333-362.

[47] Hansen, M.H. and Hurwitz, W.N. (1946) The problem of non-response in sample surveys. *J. Amer. Statist. Assoc.*, **41**, 517-529.

[48] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953) *Sample Survey Methods and Theory*. Vol. I and II. John Wiley & Sons, Inc., New York.

[49] Hansen, M.H., Hurwitz, W.N., and Gurney, M. (1946) Problems and methods of the sample survey of business. *J. Amer. Statist. Assoc.*, **41**, 173-189.

[50] Hanurav, T.V. (1967) Optimum utilization of auxiliary information: πPS sampling of two units from a stratum. *J. Roy. Statist. Soc. (B)*, **29**, 374-391.

[51] Hartley, H.O. and Ross, A. (1954) Unbiased ratio estimators. *Nature*, **174**, 270-271.

[52] Horvitz, D.G., Shah, B.V., and Simmons, W.R. (1967) The unrelated question randomized response model . *Proc. Soc. Statist. Sect., Amer. Statist. Assoc.*, 65-72.

[53] Horvitz, D.G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663-684.

[54] Hyman, H.H. (1954) *Interviewing in Social Research.* University of Chicago Press, Chicago.

[55] Jackson, C.H.N. (1933) On the true density of tsetse flies. *J. Anim. Ecol.*, **2**, 204-209.

[56] Jolly, G.M. (1965) Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, **52**, 225-247.

[57] Kendall, M.G. and Smith, B.B. (1939) *Tables of Random Sampling Numbers; Tracts for Computers.* No XXIC. Cambridge University Press.

[58] Khan, S. and Tripathi, T.P. (1967) The use of multivariate auxiliary information in double sampling. *J. Indian Statist. Assoc.*, **5,** 42-48.

[59] Kish, L. (1965) *Survey Sampling.* John Wiley & Sons, Inc., New York.

[60] Krishnamoorthy, K. and Raghavarao, D. (1993) Untruthful answering in repeated randomized response procedures. *Canadian J. Statist.*, **21**, 233-236.

[61] Lahiri, D.B. (1951) A method of sample selection providing unbiased ratio estimates. *Bull. Inter. Statist. Inst.*, **33**, 133-140.

[62] Lahiri, D.B. (1954) Technical paper on some aspects of the development of the sample design. *Sankhyā,* **14**, 264-316.

[63] Lahiri, D.B. (1963) Multi-subject sample survey system - Some thoughts based on Indian experience. *Contributions to Statistics,* 175-220. Pergamon Press, London, and Statistical Publishing Society, Calcutta.

[64] Lakshmi, D.V. and Raghavarao, D. (1992) A test for detecting untruthful answering in randomized response procedures. *J. Statist. Planning Infer.*, **31**, 387-390.

[65] Lieberman, G.J. and Owen, D.B. (1961) *Tables of the Hypergeometric Probability Distribution.* Stanford University Press.

[66] Lincoln, F.C. (1930) *Calculating Waterfowl Abundance on the Basis of Banding Returns.* Circ. U.S. Dept. Agric., No 118, May 1930.

[67] Liu, P.T., Chow, L.P., and Mosley, W.H. (1975) Use of the randomized response technique with a new randomizing device. *J. Amer. Statist. Assoc.*, **70**, 329-332.

[68] Madow, W.G. (1949) On the theory of systematic sampling II. *Ann. Math. Statist.*, **20**, 333-354.

[69] Mahalanobis, P.C. (1938) *Statistical Report on the Experimental Crop Census, 1937.* Indian Central Jute Committee.

[70] Mahalanobis, P.C. (1940) A sample survey of the acreage under jute in Bengal. *Sankhyā*, **4**, 511-530.

[71] Mahalanobis, P. C. (1952) Some aspects of the design of sample surveys. *Sankhyā*, **12**, 1-7.

[72] Mangat, N.S. (1993) Estimation of population total using an alternative estimator for RHC scheme. *Statistica*, **53**, 251-259.

[73] Mangat, N.S. (1994) An improved randomized response strategy. *J. Roy. Statist. Soc. (B)*, **56**, 93-95.

[74] Mangat, N.S. and Singh, Ravindra (1990) An alternative randomized response procedure. *Biometrika*, **77**, 439-442.

[75] Mangat, N.S. and Singh, Ravindra (1991a) An alternative approach to randomized response survey. *Statistica*, **51**, 327-332.

[76] Mangat, N.S. and Singh, Ravindra (1991b) An alternative randomized response procedure for sampling without replacement. *J. Indian Statist. Assoc.*, **29**, 127-131.

[77] Mangat, N.S., Singh, Ravindra, and Singh, Sarjinder (1992) An improved unrelated question randomized response strategy. *Calcutta Statist. Assoc. Bull.*, **42**, 277-281.

[78] Mangat, N.S., Singh, Ravindra, Singh, Sarjinder, Bellhouse, D.R., and Kashani, H.B. (1995) On efficiency of using distinct respondents in a randomized response survey. *Survey Meth.*, **21,** 21–23.

[79] Mehta, S.K., Singh, Ravindra, and Kishore, L. (1995) On optimum stratification for allocation proportional to strata totals. *J. Indian Statist. Assoc.* (Submitted).

[80] Mickey, M.R. (1959) Some finite population unbiased ratio and regression estimators. *J. Amer. Statist. Assoc.*, **54**, 594-612.

[81] Midzuno, H. (1952) On the sampling system with probability proportional to sum of sizes. *Ann. Inst. Statist. Math.*, **3**, 99-107.

[82] Moors, J.J.A. (1971) Optimization of the unrelated question randomized response model. *J. Amer. Statist. Assoc.*, **66**, 627-629.

[83] Murthy, M.N. (1957) Ordered and unordered estimators in sampling without replacement. *Sankhyā* , **18**, 379-390.

[84] Murthy, M.N. (1967) *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

[85] Murthy, M.N. and Nanjamma, N.S. (1959) Almost unbiased ratio estimates based on interpenetrating sub-sample estimates. *Sankhyā*, **21**, 381-392.

[86] Narain, R.D. (1951) On sampling without replacement with varying probabilities. *J. Indian Soc. Agric. Statist.*, **3**, 169-174.

[87] Neyman, J. (1934) On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection . *J. Roy. Statist. Soc.*, **97**, 558-606.

[88] Neyman, J. (1938) Contributions to the theory of sampling human populations. *J. Amer. Statist. Assoc.*, **33**, 101-116.

[89] Odum, E.P. and Pontin, A.J. (1961) Population density of the underground ant, *Lasius flavus,* as determined by tagging with P^{32}. *Ecology,* **42**, 186-188.

[90] Payne, S.L. (1951) *The Art of Asking Questions.* Princeton University Press, Princeton, N.J.

[91] Petersen, G.G.J. (1896) *The Yearly Immigration of Young Plaice into the Limfjord from the German Sea, etc.* Rept. Danish Biol. Statist. for 1895, **6**, 1-48.

[92] Politz, A. and Simmons, W.R. (1949) An attempt to get the "not at home" into the sample without callbacks. *J. Amer. Statist. Assoc.*, **44**, 9-31.

[93] Pollock, K.H., Hines, J.E., and Nichols, J.D. (1984) The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics,* **40**, 329-340.

[94] Quenouille, M.H. (1956) Notes on bias in estimation. *Biometrika,* **43**, 353-360.

[95] Rana, R.S. and Singh, Ravindra (1989) Note on systematic sampling with supplementary observations. *Sankhyā (B)*, **51**, 205-221.

[96] Rand Corporation (1955) *A Million Random Digits and 100000 Normal Deviates.* The Free Press, Glencoe, Illinois.

[97] Rao, C.R., Mitra, S.K., Matthai, A., and Ramamurthy, K. G. (1974) *Formulae and Tables for Statistical Work.* Statistical Publishing Society, Calcutta.

[98] Rao, J.N.K. (1965) On two simple schemes of unequal probability sampling without replacement. *J. Indian Statist. Assoc.*, **3,** 173-180.

[99] Rao, J.N.K. (1966) Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā (A)*, **28**, 47-60.

[100] Rao, J.N.K. (1969) Ratio and regression estimators. *New Developments in Survey Sampling.* N.L. Johnson and H. Smith. Jr. (eds), John Wiley & Sons, Inc., New York, 213-214.

[101] Rao, J.N.K. (1973) On double sampling for stratification and analytical surveys. *Biometrika,* **60**, 125-133.

[102] Rao, J.N.K., Hartley, H.O., and Cochran, W.G. (1962) A simple procedure of unequal probability sampling without replacement. *J. Roy. Statist. Soc. (B),* **24**, 482-491.

[103] Robson, D.S. and Regier, H.A. (1964) Sample size in Petersen mark-recapture experiments. *Trans. Amer. Fish. Soc.*, **93**, 215-226.

[104] Sampford, M.R. (1967) On sampling without replacement with unequal probabilities of selection. *Biometrika*, **54**, 499-513.

[105] Scheaffer, R.L., Mendenhall, W., and Ott, L. (1979) *Elementary Survey Sampling.* Duxbury Press, Boston, Massachusetts.

[106] Schnabel, Z.E. (1938) Estimation of the total fish population of a lake. *Amer. Math. Mon.,* **45**, 348-352.

[107] Seber, G.A.F. (1965) A note on the multiple-recapture census. *Biometrika,* **52**, 249-259.

[108] Seber, G.A.F (1973) *Estimation of Animal Abundance and Related Parameters.* Griffin, London.

[109] Sen, A.R. (1952) Present status of probability sampling and its use in estimation of farm characteristics. *Econometrica,* **20**, 130.

[110] Sen, A.R. (1953) On the estimate of the variance in sampling with varying probabilities. *J. Indian Soc. Agric. Statist.,* **5**, 119-127.

[111] Singh, D. and Chaudhary, F.S. (1989) *Theory and Analysis of Sample Survey Designs.* Wiley Eastern Limited, New Delhi.

[112] Singh, Padam and Srivastava, A.K. (1980) Sampling schemes providing unbiased regression estimators. *Biometrika,* **67**, 205-209.

[113] Singh, Ravindra (1971) Approximately optimum stratification on the auxiliary variable. *J. Amer. Statist. Assoc.,* **66**, 829-833.

[114] Singh, Ravindra (1972) A note on sampling over two occasions. *Austral. J. Statist.,* **14**, 120-122.

[115] Singh, Ravindra (1975) On optimum stratification for proportional allocation. *Sankhyā (C),* **37**, 109-115.

[116] Singh, Ravindra and Gupta., J.P. (1972) On the inbetween modification of size measure in PPSWR sampling. *J. Indian Soc. Agric. Statist.,* **24,** 49-54.

[117] Singh, Ravindra and Kishore, L. (1975) On Rao, Hartley and Cochran's method of sampling. *Sankhyā (C),* **37**, 88-94.

[118] Singh, Ravindra and Lal, M. (1978) On the construction of random groups in the RHC scheme. *Sankhyā (C),* **40**, 129-135.

[119] Singh, Ravindra, Mangat, N.S., and Singh, Sarjinder (1993) A mail survey design for sensitive character without using randomization device. *Commun. Statist.- Theor. Meth.,* **22**, 2661-2668.

[120] Singh, Ravindra and Parkash, Dev (1975) Optimum stratification for equal allocation. *Ann. Inst. Statist. Math.,* **27**, 273-280.

[121] Singh, Sarjinder and Singh, Ravindra (1992) Improved Franklin's model for randomized response sampling. *J. Indian Statist. Assoc.,* **30**, 109-122.

[122] Singh, Sarjinder and Singh, Ravindra (1993) Generalized Franklin's model for randomized response sampling. *Commun. Statist.- Theor. Meth.,* **22**, 741-755.

[123] Slonim, M.J. (1960) *Sampling in a Nutshell.* Simon & Schuster, New York.

[124] Srinath, K.P. (1971) Multiphase sampling in nonresponse problems. *J. Amer. Statist. Assoc.,* **66**, 583-586.

[125] Srivastava, S.K. (1967) An estimator using auxiliary information in sample surveys. *Calcutta Statist. Assoc. Bull.,* **16,** 121-132.

[126] Stein, C. (1945) A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.,* **16**, 243-258.

[127] Stephan, F.F., Deming, W.E., and Hansen, M.H. (1940) The sampling procedure of the 1940 population census. *J. Amer. Statist. Assoc.,* **35**, 615-630.

[128] Sukhatme, P. V. (1953) *Sampling Theory of Surveys with Applications.* Iowa State Univ. Press, Ames, Iowa, and Indian Soc. Agric. Statist., New Delhi.

[129] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., and Asok, C. (1984). *Sampling Theory of Surveys with Applications.* Iowa State Univ. Press, Ames, Iowa, and Indian Soc. Agric. Statist., New Delhi.

[130] Swain, A.K.P.C. (1964) The use of systematic sampling in ratio estimates. *J. Indian Statist. Assoc.,* **2**, 160-164.

[131] Tin, M. (1965) Comparison of some ratio estimators. *J. Amer. Statist. Assoc.,* **60**, 294-307.

[132] Tippett, L.H.C. (1927) *Random Sampling Numbers; Tracts for Computers,* No. XV. Cambridge University Press.

[133] Tracy, Derrick S. and Osahan, S.S. (1994) Estimation in overlapping clusters with unknown population size. *Survey Meth.*, **20**, 53-57.

[134] Tracy, Derrick S. and Mangat, N.S. (1995a) Respondent's privacy hazards in Moors' randomized response model- a remedial strategy. Inter. *J. Math. Statist. Sci.,* **4**, 1-10.

[135] Tracy, Derrick S. and Mangat, N.S. (1995b) On respondent's jeopardy in two alternate questions randomized response model. *J. Statist. Planning Infer.,* (In press).

[136] Warner, S.L. (1965) Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.,* **60**, 63-69.

[137] Watson, D.J. (1937) The estimation of leaf areas. *J. Agric. Sci.,* **27**, 474-483.

[138] Williams, W.H. (1963) The precision of some unbiased regression estimators. *Biometrics,* **19**, 352-361.

[139] Yates, F. (1948) Systematic sampling. *Phil. Trans. Roy. Soc. (A),* **241**, 345-377.

[140] Yates, F. (1960) *Sampling Methods for Censuses and Surveys.* Charles Griffin & Company Limited, London.

[141] Yates, F. and Grundy, P.M. (1953) Selections without replacement from within strata with probability proportional to size. *J. Roy. Statist. Soc. (B)*, **15**, 253-261.

[142] Zarkovich, S.S. (1961) *Sampling Methods and Censuses.* Vol.I. Food and Agricultural Organization of the United Nations, Rome.

[143] Zarkovich, S.S. (1966) *Quality of Statistical Data.* Food and Agricultural Organization of the United Nations, Rome.

[144] Zarkovich, S.S. and Krane, J. (1965) Some efficient uses of compact cluster sampling. *Proc. Inter. Statist. Inst.,* Belgrade Session.

[145] Zinger, A. (1963) Estimation de variance avec echantillonnage systematique. *Revue de statistique Appliquee,* **11**, 89-97.

[146] Zinger, A. (1964) Systematic sampling in forestry. *Biometrics,* **20**, 553-565.

Author Index

BOOKS/ MONOGRAPHS

Chakravarti, I.M., Laha, R.G., and Roy, J. (1967), 373
Cochran, W.G., (1977), 26, 50, 239, 242, 243, 283, 331, 373

Deming, W.E. (1960), 11, 374
Des Raj (1968), 3, 11, 243, 374

Fisher, R.A. and Yates, F. (1938), 31, 374
Fisz, M. (1963), 25, 375

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953), 139, 375
Hyman, H.H. (1954), 3, 376

Kendall, M.G. and Smith, B.B (1939), 31, 376
Kish, L.(1965), 376

Lahiri, D.B. (1963), 6,376
Lieberman, G.J and Owen, D.B. (1961), 328, 376
Lincoln, F.C. (1930), 314, 376

Mahalanobis, P.C. (1938), 67, 376
Murthy, M.N. (1967), 3, 48, 67, 278, 283, 377

Payne, S.L. (1951), 3, 378
Petersen, G.G.J. (1896), 314, 315, 317, 378

Rand Corporation (1955), 31, 378
Rao, C.R., Mitra, S.K., Matthai, A., and Ramamurthy, K.G. (1974), 31, 378

Scheaffer, R.L., Mendenhall, W., and Ott, L. (1979), 379
Seber, G. A. F. (1973), 328, 379
Singh, D. and Chaudhary, F.S. (1989), 67, 243, 379
Slonim, M.J. (1960), 11, 380
Sukhatme, P.V. (1953), 78, 380
Sukhatme, P. V., Sukhatme, B.V., Sukhatme, S., and Asok, C. (1984), 50, 67, 84, 138, 139, 278, 283, 307, 331, 380,

Tippett, L.H.C. (1927), 31, 380

Yates, F. (1960), 197, 380

Zarkovich, S.S. (1961), 6, 381
Zarkovich, S.S (1966), 331, 381

RESEARCH PAPERS

Abernathy, J.R., 347, 348, 350, 373, 375
Abul-Ela, A.L.A., 341, 345, 346, 350, 353, 373, 375
Adams, L., 317, 373
Adhvaryu, D., 186, 375
Aoyama, H., 139, 373

Badaloni, M., 328, 373
Bansal, M.L., 98, 373
Bailey, N.T.J., 314, 317, 318, 320, 373
Beale, E.M.L., 185, 373
Bellhouse, D.R., 344, 377
Best, D.J., 56, 373
Bowley, A.L., 111, 373
Brewer, K.R.W., 97, 373

Chapman, D.G., 314, 315, 320, 322, 373
Chow, L.P., 357, 376
Cochran, W.G., 91, 283, 373, 378
Cox, D.R., 26, 374

Dalenius, T., 133, 135 139, 374
Das, A.C., 78, 374
Deming, W.E., 146, 337, 374, 380
Des Raj, 48, 78, 217, 374
Dey, A., 97, 374
Diana, G., 185, 374
Durbin, J., 97, 336, 374

Eichhorn, B.H., 357, 374
El-Badry, M.A., 336, 374

Finney, D.J., 56, 374
Folsom, R.E., 347, 375
Franklin, L.A., 344, 375

Ghangurde, P.D., 242, 375
Ghosh, B., 216, 375
Goodman, R., 97, 139, 375
Greenberg, B.G., 341, 345, 346, 347, 348, 350, 353, 373, 375
Grundy, P.M., 84, 381
Gupta, J.P., 98, 379
Gupta, P.C., 186, 375
Gurney, M., 139, 179, 374, 375

Haldane, J.B.S., 56, 375
Hansen, M.H., 67, 146, 179, 283, 332, 375, 380
Hanurav, T.V., 97, 375
Hartley, H.O., 91, 184, 375, 378
Hayre, L.S., 357, 374
Hines, J.E., 328, 378
Hodges, J.L.Jr., 133, 135, 139, 374
Horvitz, D.G., 78, 84, 341, 344, 345, 346, 347, 348, 350, 353, 373, 375, 376
Hurwitz, W.N., 67, 179, 283, 332, 375

Jackson, C.H.N., 314, 376
Jolly, G.M., 328, 376

Kashani, H.B., 344, 377
Khamis, S.H., 48, 374
Khan, S., 242, 376
Kish, L., 97, 139, 375
Kishore, L., 98, 140, 377, 379
Krane, J., 278, 381
Krishnamoorthy, K., 341, 376
Kuebler, R.R., 350, 375

Lahiri, D.B., 69, 71, 146, 283, 376
Lakshmi, D.V.,341, 376
Lal, M., 98, 379
Liu, P.T., 357, 376

Madow, W.G., 97, 376
Mahalanobis, P.C., 139, 283, 377
Mangat, N.S., 97, 342, 344, 347, 348, 377, 379, 380
Mehta, S.K., 140, 377
Mickey, M.R., 184, 216, 377
Midzuno, H., 86, 184, 377
Moors, J.J.A., 347, 377
Mosley, W.H., 357, 376
Murthy, M.N., 78, 82, 185, 377

Nanjamma, N.S., 185, 377
Narain, R.D., 97, 377

Neyman, J., 117, 243, 377, 378
Nichols, J.D., 328, 378

Odum, E.P., 319, 378
Osahan, S.S., 278, 380

Parkash, Dev, 140, 379
Politz, A., 337, 378
Pollock, K.H., 328, 378
Pontin A.J., 319, 378

Quenouille, M.H., 185, 378

Raghavarao, D., 341, 376
Rana, R.S., 153, 378
Rao, J.N.K., 91, 97, 98, 216, 242, 243, 375, 378
Regier, H.A., 324, 325, 378
Robson, D.S., 324, 325, 378
Ross, A., 184, 375

Sampford, M.R., 97, 378
Schnabel, Z.E., 328, 379
Seber, G.A.F., 328, 379
Sen, A.R., 84, 86, 184, 379
Shah, B.V., 344, 345, 376
Simmons, W.R., 337, 344, 345, 346, 350 353, 375, 376, 378
Singh, Padam, 216, 379
Singh, Ravindra, 98, 135, 140, 143, 153, 242, 342, 344, 348, 373, 377, 378, 379
Singh, Sarjinder, 344, 348, 377, 379
Srinath, K.P., 243, 380
Srivastava, A.K., 97, 216, 374, 379
Srivastava, S.K., 185, 380
Stein, C., 26, 380
Stephan, F.F., 146, 380
Swain, A.K.P.C., 185, 380

Thompson, D.J., 78, 84, 376
Tin, M., 185, 380
Tracy, Derrick S., 278, 347, 380
Tripathi, T.P., 242, 376

Warner, S.L., 340, 341, 343, 344, 380
Watson, D.J., 197, 380
Williams, W.H., 216, 380

Yates, F., 84, 160, 380, 381

Zarkovich, S.S., 278, 381
Zinger, A., 153, 381

Subject Index

Accuracy, 23
Advantages of sample survey, 6
Allocation of sample, 108
Animal populations, *see* Mobile populations

Bailey's estimator, 318, 321
Balanced systematic sampling, 161
Bias, 20
  effect of bias, 25
Biased estimator, definition, 20
Binomial model, 315, 317
Bound on the error of estimation, 26

Call-backs technique, 337
Capture-recapture method, 314
Census, 5
Chapman's estimator, 315
Circular systematic sampling, 148
Cluster sampling, definition, 248
  estimation of mean, 250
  estimation of proportion, 269
  estimation of total, 258
  notations, 249
  relative efficiency, 264
  sample size, 266, 273
  unequal probability selection, 275
Clusters, 248
Collapsed strata method, 139
Combined product estimator, 190
Combined ratio estimator, 181
Combined regression estimator, 211
Complete enumeration, 5
Confidence bound, 26
Confidence coefficient, 24
Confidence interval, 23, 24, 46
Confidence limits, 24, 46
  lower limit, 24, 46
  upper limit, 24, 46
Consistency, 23
Consistent estimator, 23
Construction of strata, 132
Controlled selection, 139
Convenience sampling, 7

Cost function
  for mobile populations, 325
  in nonresponse errors, 335
  in stratified sampling, 108, 117
  in two-phase sampling, 231
Covariance, definition, 15
Cumulative cube root method, 135
Cumulative square root method, 133
Cumulative total method, 68

Data collection, methods of, 2, 10
Deep stratification, 139
Difference estimator, definition, 197
  estimation of mean, 198
  multivariate difference estimator, 217
Distinct units based estimator, 48
Double sampling, *see* Two-phase sampling

Effect of bias, 25
Element, 4
End correction, 160
Enumeration, complete, 5
Equal allocation, 108
Equal probability sampling, *see* Simple random sampling
Estimate, definition, 17
Estimate of standard error, 22
Estimation variable, 102
Estimator, definition, 17
  biased, 20
  consistent, 23
  unbiased, 19
Execution of sample survey, 9
Expectation, 14

Field work, organization of, 11
Final sample, 221
Finite population, 4
Finite population correction, 40
First phase sample, 221
First stage units, 283
Frame, 5, 10

Hansen and Hurwitz technique, 332
Horvitz and Thompson estimator, 84
Hypergeometric model, 315

Inclusion probability proportional
  to size (IPPS or πPS) procedure, 84
Infinite population, 4
Initial sample, 221
Interpenetrating subsamples, 153
Interval estimate, 24
Interview, 2, 3
  by enumerators, 3
  personal, 2
  through telephone, 3
Inverse sampling, definition, 56, 315, 320
  binomial, 56, 322
  hypergeometric, 320

Judgement sampling, 7

Lahiri's method, 69
Linear regression estimator, *see*
  Regression estimator
Linear trend, 146, 160
Linear systematic sampling, 145
Lottery method, 31
Lower confidence limit, 24, 46

Mail inquiry, 2
Mail survey, *see* Mail inquiry
Mangat and Singh's two-stage model, 342
Mean square error, definition, 22
Measures of error, 21
Midzuno system of sampling, *see*
  Sen-Midzuno method
Mobile populations, 314
  assumptions, 314
  Bailey's estimator, 318, 321
  binomial model, 315, 317
  capture-recapture method, 314
  Chapman's estimator, 315
  cost function, 325
  direct sampling, *see* Binomial model
  hypergeometric model, 315
  inverse sampling, 315
  multiple markings, 328
  negative binomial, 320, 322
  negative hypergeometric, 320
  Petersen estimator, 317, 322
  sample size, 324
Modified systematic sampling, 160
Multiphase sampling, *see* Two-phase sampling
Multiple markings, 328
Multiple sampling, 237
Multiple stratification, 139
Multistage sampling, *see* Two-stage sampling
Multivariate difference estimator, 217

Negative binomial, 56, 320, 322
Negative hypergeometric, 56, 320
Neyman allocation, 117, 121
Nonprobability sampling, 7
Nonresponse error, 331, 340
Nonsampling errors, 6, 27, 331
  cost function, 335
  estimation of mean, 332, 337, 350
  estimation of proportion, 340, 343, 345
Normal distribution, 19, 24, 364
Notations, 14
  for cluster sampling, 249
  for stratified sampling, 104
  for two-stage sampling, 284

Objectives of sample survey, 9
Optimum allocation, 114
Ordered estimator, 78
Overlapping clusters, 278

Parameter, 17
Periodic population, 146
Permissible error, 26
Personal interview, 2
Petersen estimator, 317, 322
Pilot survey, 10
Planning of sample survey, 9
Point estimate, 17
Politz and Simmons' technique, 337
Population, definition, 4
  mean, 17
  mean square error, 39
  parameter, 17
  sampled, 9
  size, 6
  standard deviation, 17
  variance, 17
Post-stratification, 136
Precision, 23
Preliminaries, 14
Primary data, definition, 2
  methods of collection, 2, 10
Primary stage units, 283
Primary strata, 139
Probability proportional to size sampling,
  definition, 67

estimation of mean, 70
estimation of total, 70, 78, 82, 84, 92,
in cluster sampling, 275
in stratified sampling, 138
in two-phase sampling, 233
in two-stage sampling, 304
relative efficiency, 72, 95
sample size, 76
with replacement sampling, 70
without replacement sampling, 77
Probability sampling, 7
Product estimator, 185
combined product estimator, 190
efficiency, 186
estimation of mean, 185
estimation of total, 186
sample size, 189
separate product estimator, 190
Proportion, 53
Proportional allocation, 111
Purposive sampling, 7

Qualitative variable, sensitive, 340
Quantitative variable, sensitive, 350
Questionnaire, framing of, 3
Quota sampling, 7

Random group method, 91
Random number tables, 31, 365
Random numbers, definition, 31
use in selecting PPS sample, 67, 69
use in selecting simple random sample, 31
Random start, 145
Random variable, 17
Randomized response technique, 340
estimation of mean, 350
estimation of proportion, 340, 343, 345
Rao, Hartley, Cochran's scheme, 91
relative efficiency, 95
Ratio estimation, need for, 165
Ratio estimator, definition, 165
combined ratio estimator, 178, 179, 181
estimation of mean, 169, 179, 182
estimation of ratio, 166
estimation of total, 170
sample size, 175
separate ratio estimator, 178, 179
unbiased ratio type estimator, 184
Regression estimator, definition, 201
combined regression estimator, 211
estimation of mean, 201, 207, 212
sample size, 204
separate regression estimator, 206
unbiased regression estimator, 216
Relative bias, definition, 20
Relative efficiency, definition, 22
of cluster sampling, 264
of PPSWR sampling, 72
of RHC scheme, 95
of stratified sampling, 123
Report writing, 11
Revised probabilities of selection, 87
Root mean square error, 22

Sample, definition, 5
need for a sample, 5
Sample mean, 17
Sample mean square, 17
Sample size, 6, 26
determination, 26
Sample survey, advantages of, 6
Sampling, definition, 7
Sampling distribution, definition, 18
of mean, 18, 37, 41
Sampling errors, 6, 27
Sampling fraction, 6
Sampling frame, 5, 10
Sampling interval, 145
Sampling on two occasions, 237
Sampling procedures, 6
Sampling units, 5, 10
Sampling variance, definition, 21
Sampling variance of mean, 33
Sampling with varying probabilities, *see* Probability proportional to size sampling
Schedule, 2, 3
Second phase sample, 221
Second stage units, 283
Secondary data, definition, 2
Secondary units, 283
Sen-Midzuno method, 86
Sensitive questions, 340
Sensitive variable, qualitative, 340
Sensitive variable, quantitative, 350
Separate product estimator, 190
Separate ratio estimator, 179
Separate regression estimator, 206
Simple random sample, definition, 30
Simple random sample, selection of, 30
direct approach, 31
quotient approach, 31
remainder approach, 32
Simple random sampling, definition, 30
estimation of mean, 33, 40, 48, 59
estimation of proportion, 53, 56, 61
estimation of total, 45, 48
sample size, 50, 55
with replacement sampling, 30, 33

without replacement sampling, 30, 39
Standard deviation, 17
Standard error, definition, 21
estimate, 22
Statistic, definition, 17
Statistical data, need for, 1
Strata, 102
Stratification, 102
Stratification variable, 102
Stratified random sampling, definition, 102
Stratified sampling, definition, 102
advantages, 103
allocation of sample size, 108
construction of strata, 132
controlled selection procedure, 139
cost function, 108, 117
deep stratification, 139
estimation of mean, 104
estimation of proportion, 129
estimation of total, 106
multiple stratification, 139
notations, 104
post stratification, 136
principles of stratification, 103
relative efficiency, 123
with varying probabilities, 138
Subpopulation, 58
estimation of mean, 59
estimation of proportion, 61
Subunits, 283
Successive sampling, definition, 237
Systematic sampling, definition, 145
estimation of mean, 149, 153
estimation of proportion, 158
estimation of total, 149, 153
in periodic population, 146
in population with linear trend, 160
interpenetrating subsamples, 153
sample size, 156
selection of sample, 145

Target population, 9
Telephone interview, 3
Tolerable error, 26
Two-phase sampling, definition, 222
cost function, 231
multiphase sampling, 222
need for two-phase sampling, 221
product method of estimating mean, 226
ratio method of estimating mean, 223
regression method of estimating mean 229
sample size, 231
varying probability sampling, 233
Two-stage sampling, definition, 283
estimation of mean, 284, 304
estimation of proportion, 296
estimation of total, 294
multistage sampling, definition, 283
notations, 284
use of PPS sampling, 304
Types of data, 1

U-model, 344
Unbiased estimator, definition, 19
Unbiased ratio type estimator, 184
Unbiased regression estimator, 216
Unequal probability sampling, *see*
Probability proportional to size sampling
Universe, 4
Unordered estimator, 82
Unrelated question RR model, 344, 345
Upper confidence limit, 24, 46

Variance, definition, 15
Variance of linear functions, 16
Varying probability sampling, *see*
Probability proportional to size sampling

Warner's randomized response model, 340
Wildlife populations, *see* Mobile populations
With replacement sampling, 7
Without call-backs procedure, *see*
Politz and Simmons' technique
Without replacement sampling, 7, 8

Yates end correction, 160

Kluwer Texts in the Mathematical Sciences

1. A.A. Harms and D.R. Wyman: *Mathematics and Physics of Neutron Radiography.* 1986 ISBN 90-277-2191-2
2. H.A. Mavromatis: *Exercises in Quantum Mechanics.* A Collection of Illustrative Problems and Their Solutions. 1987 ISBN 90-277-2288-9
3. V.I. Kukulin, V.M. Krasnopol'sky and J. Horácek: *Theory of Resonances.* Principles and Applications. 1989 ISBN 90-277-2364-8
4. M. Anderson and Todd Feil: *Lattice-Ordered Groups.* An Introduction. 1988 ISBN 90-277-2643-4
5. J. Avery: *Hyperspherical Harmonics.* Applications in Quantum Theory. 1989 ISBN 0-7923-0165-X
6. H.A. Mavromatis: *Exercises in Quantum Mechanics.* A Collection of Illustrative Problems and Their Solutions. Second Revised Edition. 1992 ISBN 0-7923-1557-X
7. G. Micula and P. Pavel: *Differential and Integral Equations through Practical Problems and Exercises.* 1992 ISBN 0-7923-1890-0
8. W.S. Anglin: *The Queen of Mathematics.* An Introduction to Number Theory. 1995 ISBN 0-7923-3287-3
9. Y.G. Borisovich, N.M. Bliznyakov, T.N. Fomenko and Y.A. Izrailevich: *Introduction to Differential and Algebraic Topology.* 1995 ISBN 0-7923-3499-X
10. J. Schmeelk, D. Takači and A. Takači: *Elementary Analysis through Examples and Exercises.* 1995 ISBN 0-7923-3597-X
11. J.S. Golan: *Foundations of Linear Algebra.* 1995 ISBN 0-7923-3614-3
12. S.S. Kutateladze: *Fundamentals of Functional Analysis.* 1996 ISBN 0-7923-3898-7
13. R. Lavendhomme: *Basic Concepts of Synthetic Differential Geometry.* 1996 ISBN 0-7923-3941-X
14. G.P. Gavrilov and A.A. Sapozhenko: *Problems and Exercises in Discrete Mathematics.* 1996 ISBN 0-7923-4036-1
15. R. Singh and N. Singh Mangat: *Elements of Survey Sampling.* 1996. ISBN 0-7923-4045-0

KLUWER ACADEMIC PUBLISHERS – DORDRECHT / BOSTON / LONDON