Survey Methods and Sampling Theory

STA 324

University of Ibadan Distance Learning Centre
Open and Distance Learning Course Series Development

ISBN: 978-021-556-5

General Editor: Prof. Bayo Okunade

University of Ibadan Distance Learning Centre
University of Ibadan,
Nigeria

Telex: 31128NG

Tel: +234 (80775935727)
E-mail: ssu@dlc.ui.edu.ng
Website: www.dlc.ui.edu.ng

Vice-Chancellor's Message

The Distance Learning Centre is building on a solid tradition of over two decades of service in the provision of External Studies Programme and now Distance Learning Education in Nigeria and beyond. The Distance Learning mode to which we are committed is providing access to many deserving Nigerians in having access to higher education especially those who by the nature of their engagement do not have the luxury of full time education. Recently, it is contributing in no small measure to providing places for teeming Nigerian youths who for one reason or the other could not get admission into the conventional universities.

These course materials have been written by writers specially trained in ODL course delivery. The writers have made great efforts to provide up to date information, knowledge and skills in the different disciplines and ensure that the materials are user-friendly.

In addition to provision of course materials in print and e-format, a lot of Information Technology input has also gone into the deployment of course materials. Most of them can be downloaded from the DLC website and are available in audio format which you can also download into your mobile phones, IPod, MP3 among other devices to allow you listen to the audio study sessions. Some of the study session materials have been scripted and are being broadcast on the university's Diamond Radio FM 101.1, while others have been delivered and captured in audio-visual format in a classroom environment for use by our students. Detailed information on availability and access is available on the website. We will continue in our efforts to provide and review course materials for our courses.

However, for you to take advantage of these formats, you will need to improve on your I.T. skills and develop requisite distance learning Culture. It is well known that, for efficient and effective provision of Distance learning education, availability of appropriate and relevant course materials is a *sine qua non*. So also, is the availability of multiple plat form for the convenience of our students. It is in fulfilment of this, that series of course materials are being written to enable our students study at their own pace and convenience.

It is our hope that you will put these course materials to the best use.

Prof. Abel Idowu Olayinka

Vice-Chancellor

Foreword

As part of its vision of providing education for "Liberty and Development" for Nigerians and the International Community, the University of Ibadan, Distance Learning Centre has recently embarked on a vigorous repositioning agenda which aimed at embracing a holistic and all encompassing approach to the delivery of its Open Distance Learning (ODL) programmes. Thus we are committed to global best practices in distance learning provision. Apart from providing an efficient administrative and academic support for our students, we are committed to providing educational resource materials for the use of our students. We are convinced that, without an up-to-date, learner-friendly and distance learning compliant course materials, there cannot be any basis to lay claim to being a provider of distance learning education. Indeed, availability of appropriate course materials in multiple formats is the hub of any distance learning provision worldwide.

In view of the above, we are vigorously pursuing as a matter of priority, the provision of credible, learner-friendly and interactive course materials for all our courses. We commissioned the authoring of, and review of course materials to teams of experts and their outputs were subjected to rigorous peer review to ensure standard. The approach not only emphasizes cognitive knowledge, but also skills and humane values which are at the core of education, even in an ICT age.

The development of the materials which is on-going also had input from experienced editors and illustrators who have ensured that they are accurate, current and learner-friendly. They are specially written with distance learners in mind. This is very important because, distance learning involves non-residential students who can often feel isolated from the community of learners.

It is important to note that, for a distance learner to excel there is the need to source and read relevant materials apart from this course material. Therefore, adequate supplementary reading materials as well as other information sources are suggested in the course materials.

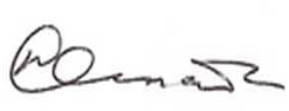
Apart from the responsibility for you to read this course material with others, you are also advised to seek assistance from your course facilitators especially academic advisors during your study even before the interactive session which is by design for revision. Your academic advisors will assist you using convenient technology including Google Hang Out, You Tube, Talk Fusion, etc. but you have to take advantage of these. It is also going to be of immense advantage if you complete assignments as at when due so as to have necessary feedbacks as a guide.

The implication of the above is that, a distance learner has a responsibility to develop requisite distance learning culture which includes diligent and disciplined self-study, seeking available administrative and academic support and acquisition of basic information technology skills. This is why you are encouraged to develop your computer skills by availing yourself the opportunity of training that the Centre's provide and put these into use.

In conclusion, it is envisaged that the course materials would also be useful for the regular students of tertiary institutions in Nigeria who are faced with a dearth of high quality textbooks. We are therefore, delighted to present these titles to both our distance learning students and the university's regular students. We are confident that the materials will be an invaluable resource to all.

We would like to thank all our authors, reviewers and production staff for the high quality of work.

Best wishes.



Professor Bayo Okunade

Director

Course Development Team

Content Authoring	Sodipo A.A
	Onyeze V.C
Content Editor	Prof. Remi Raji-Oyelade
Production Editor	Ogundele Olumuyiwa Caleb
Learning Design/Assessment Authoring	SkulPortal Technology
Managing Editor	Ogunmefun Oladele Abiodun
General Editor	Prof. Bayo Okunade

Table of Contents

Study Session 1: Introduction to survey methods and sampling theory 13

Introduction 13

Learning Outcomes for Study Session 1 13

1.1 Basic Definitions and concepts 13

1.2 Sample Survey 15

1.2.1 Advantages of Sample Survey 16

1.2.2 Disadvantages of Sample Survey 16

1.3 Planning and execution of sample surveys 17

1.4 Activities that could involve sample survey 21

Summary for Study Session1 23

Self-Assessment Questions (SAQs) for Study Session 1 24

SAQ 1.1 (Testing Learning outcomes 1.1) 24

SAQ 1.2 (Testing Learning outcomes 1.2) 24

SAQ 1.3 (Testing Learning outcomes 1.3) 24

SAQ 1.4 (Testing Learning outcomes 1.4) 24

Notes on SAQ 24

SAQ 1.1 24

SAQ 1.2 24

SAQ 1.3 25

SAQ1.4 25

References 26

Study Session 2: Probability and non-Probability sampling 27

Introduction 27

Learning Outcomes for Study Session 2 27

2.1 Probability Sampling (Random) 27

2.2.1 Simple Random Sampling (SRS) 28
2.2.2 Selection Probability 30
2.2.3 Properties of the Estimates (SRS) 30
2.2.4 Estimation of Population Mean 31
2.2.5 Estimation of Population Total 33
2.2.6 Estimation of the Standard Error from a Sample 34
2.2.7 The Finite Population Correction 35
2.2 Non-Random Sampling (non probability) 38
Summary for Study Session 2 40
Self-Assessment Questions (SAQs) for Study Session 2 40
SAQ 2.1 41
SAQ 2.2 41
Notes on SAQ 41
SAQ 2.1 41
SAQ 2.2 41
References 41
Study Session 3: Estimation of Population Proportion 42
Introduction 42
Learning outcomes for study session 3 42
3.1 Estimation of population proportion 42
3.1.1 Estimation of N_c 43
3.2 Confidence Limit 45
Summary for study sessionn3 47
Self-Assessment Questions (SAQs) for Study Session 3 48
SAQ 3.1 48
Notes on SAQ 48

SAQ 3.1 48
References 49
Study Session 4: Systematic Sampling 50
Introduction 50
Learning Outcomes for study session 4 50
4.1 Systematic Sampling 50
4.2 Estimation in Systematic Sampling 52
4.3 Determination of Sample Size 54
Summary for Study Session 4 55
Self-Assessment Questions (SAQs) for Study Session 4 56
SAQ 4.1 56
SAQ 4.2 56
Notes on SAQ 56
SAQ 4.1 56
SAQ 4.2 56
References 57
Study Session 5: Stratified Random Sampling 58
Introduction 58
Learning outcomes for Study Session 5 59
5.1 Stratified Random Sampling 59
5.2 Reasons for Stratification 61
5.3 Estimation in Stratified Sampling 62
5.4 Estimation of Population Mean 63
5.5 Estimation of Population Proportion 64
5.6 Allocation of Sample Size to Strata 67
5.5.1. Proportional Allocation: 68

5.5.2 X-Proportional Allocation 69
5.5.3 Optimum Allocation 70
5.5.4 Equal Allocation 71
Summary for Study Session 5 72
Self-Assessment Questions (SAQs) for Study Session 5 72
SAQ 5.1 72
SAQ 5.2 72
Notes on SAQ 72
SAQ 5.1 72
SAQ 5.2 72
References 73
Study Session 6: Ratio and Regression Estimation 74
Introduction 74
Learning Outcomes for Study Session 6 74
6.1 Ratio Estimation 74
6.1.1 Definitions and Notations 75
6.1.2 Approximate Variance of Ratio Estimator 76
6.1.3 Comparison with the simple average 81
6.2 Regression Estimation 82
6.2.1 Regression Estimation of Population Mean 82
Summary for study session 6 85
Self-Assessment Questions (SAQs) for Study Session 6 85
SAQ 6.1 85
SAQ 6.2 85
Notes on SAQ 86
SAQ 6.1 86

SAQ 6.2 86
References 86
Study Session 7:Non-Sampling Errors 87
Introduction 87
Learning outcomes for study Session 7 87
7.1. Non-Sampling errors 87
7.1.1 The Planning Stage: 88
7.1.2 The Execution Stage: 89
7.1.3 Analysis Stage: 89
7.2 Types of Non-Sampling Error 89
7.2.1 Response Error 90
7.3 Sources of Response Error 90
7.3.1 The interviewer 91
7.3.2 Instrument 91
7.3.3 Respondent 92
7.4 Non-Response Error 93
Summary for study session 7 95
Self-Assessment Questions (SAQs) for Study Session 7 96
SAQ 7.1 (Testing Learning outcomes 7.1) 96
SAQ 7.2 (Testing Learning outcomes 7.2) 96
SAQ 7.3 (Testing Learning outcomes 7.3) 96
SAQ 7.4 (Testing Learning outcomes 7.4) 96
Notes on SAQ 96
SAQ 7.1 96
SAQ 7.2 97
SAQ 7.3 97

SAQ 7.4.....97

References.....97

Study Session 1: Introduction to survey methods and sampling theory

Introduction

Sampling is a scientific method of selecting and using a representative part (Sample) of a whole to seek the truth about the whole. Sampling is used extensively, consciously or unconsciously in everyday life to obtain the required information or to carry out a course of action.

By using a few millilitres of a patient's blood sample, a medical doctor obtains the quantity of malaria parasites in the blood system. Also, a market researcher makes use of a fraction of consumers of a product to gauge the acceptability of the product.

Sample survey in its simplest terms deals with collection of information to satisfy human everyday needs. Such needs could be in agriculture, population and labour etc.

Learning Outcomes for Study Session 1

At the end of this study, you should be able to:

1.1 Define some terms and concepts
1.2 Discuss sample surveys
1.3 Explain the planning and execution of sample survey
1.4 Identify some activities that could involve sample survey

1.1 Basic Definitions and concepts

There are some basic definitions and concepts you need to understand to facilitate a robust understanding of this study. Here are some.

Sample: This is finite sequence of elements drawn from the population,

U denoted by $S = \left\{U_{i1}, U_{i2}, \cdots U_{i_{n(S)}}\right\}$ i_j and denotes the unit drawn at the j^{th} draw, $n(s)$. This is called the sample size.

Sampling is a scientific method of selecting and using a representative part (Sample) of a whole to seek the truth about the whole. Sampling is used extensively, consciously or unconsciously in everyday life to obtain the required information or to carry out a course of action.

Sampling Units: These are the elementary units of a population from which a sample is selected.

Sampling Scheme: This is a method of selecting a sample from a population, e.g. Simple random sampling, systematic sampling, etc.

Sampling Design: This is defined as the collection of all possible samples together with their probability of selection.

Suppose $S = \{s\}$: a collection of all possible samples from a population. Let $P = \left\{P_{(s)}\right\}$ be the probability measure defined on S such that for all $s \in \mathrm{S}$, you will have $P(s) \geq 0, \sum P(s) = 1$. $D = \{S, P\}$ This is called the sampling design.

Elementary Unit or Unit: This is an element or a group of elements from which the required information is obtained, e.g. household, farm, person, etc. A reporting unit is the unit that actually supplies the required information, e.g. a head of a household may be a reporting unit in a household survey.

Population or Universe: This is a collection of units having some properties/characteristics in common. A population is finite if the number of units making up the population is finite; and infinite if the number of units in the population is infinite.

Population Parameter or Parameter: This is a function, $\phi\left(Y_1, \mathrm{Y}_2, \ldots \mathrm{Y}_{\mathrm{N}}\right)$,defined on the population values. Y_i is the value of the characteristic of interest for the i^{th} $\left(i = 1, 2, \ldots \mathrm{N}\right)$ population unit? Example of a parameter is the population mean, total, proportion or ratio.

Estimator: This is any real valued function $\hat{\phi}\left(\mathrm{y}_1, \mathrm{y}_2 \ldots y_{n_{(s)}}\right)$ defined on the parameter ϕ. The value of an estimator for a given sample is called an estimate.

Sampling Strategy: The pair, sampling design and estimator, $T\left(D,\hat{\phi}\right)$ is called a sampling strategy.

Accuracy: This is a measure of closeness of an estimate to its parameter. This is judged by MSE (the expected value of the squared error loss function) and hence includes the effect of bias.

Precision: This is a measure of how close an estimate is to its average value over all possible samples. This is judged by the sampling variance and, therefore, excludes the effect of bias.

Sampling error: This is the error that occurs as a result of using a sample to make inference about the population. The actual measure of this error is the sampling variance. The square root of the sampling variance is the sampling error.

Enumeration Area (EA): Is a small area, generally used as a sampling unit, with well-defined and identifiable foundations carved out of larger area of land.

In-Text Question

_____ is defined as the collection of all possible samples together with their probability of selection.

A. Precision
B. Sampling Strategy
C. Sample Design
D. Estimator

In-Text Answer

D. Sample Design

1.2 Sample Survey

A sample survey is defined as the collection and examination of data from a sample in order to make inferences about the whole. This is contrary to a census, which is a complete enumeration or survey of the whole units of enquiring.

Hence, sample survey theory deals with the process of sample selection, data collection, estimation of the population characteristics using the sample data to collect and determining

the accuracy of the estimates. A population characteristic is defined as any quantity relating to the population.

Information on a population may be collected in two ways. Either every unit in the population is enumerated (called complete enumeration, or census) or enumeration is limited to only a part or a sample selected from the population (**called sample enumeration or sample survey**).

1.2.1 Advantages of Sample Survey

1. In destructive investigations, such as determination of the mean, Life-time of electric bulbs or the shelf life of some perishable items, it is better to use sample survey rather than complete examination.
2. Sample survey saves time, labour and cost, especially when these resources are limited, than a census where a large number of these resources is needed.
3. Data can be collected and analysed quickly from a sample than from the entire population.
4. There is a greater and more efficient supervision of field staff in a sample survey resulting in the collection of more reliable data.
5. Sample survey makes use of better qualified staff and specialized equipment.
6. It has greater subject coverage and less observational error than census.

1.2.2 Disadvantages of Sample Survey

1. Sample survey is not appropriate when information is required for every unit of enquiry.

2. It is less accurate in small area classification where data are needed for each subdivision of the population.

In-Text Question

Sample survey saves time, labour and cost, especially when these resources are limited. True or False

In-Text Answer

True

1.3 Planning and execution of sample surveys

The following are some of the basic steps involved in planning and execution of large-scale sample survey:

1. Objectives of the Survey: The objectives of the survey must be stated in clear, concrete and concise terms. The statement of the objective should include the reason for the survey, questions to be covered, level of accuracy required and the expected results. Failure to state the objective of the survey in a precise form will undermine its ultimate value; in the end it may be found that the results are not what were really wanted.

2. Definition of the Population to be studied (Target Population): The objectives of the survey should define the population the survey is intended to cover. But practical difficulties in handling certain segments of the population may point to their elimination from the scope of the survey.

E.g. transient population may be difficult to cover in a population survey. The target population would generally be different from the population actually sampled. Sometimes, information is collected in a different manner from the omitted sector.

3. Frame: In order to cover the population decided upon, there should be some list s, map or other acceptable material which serves as a guide to the universe to be covered.

A frame is, therefore, a list or map of the population to be covered together with its identification scheme from which the sample is selected, e.g. list of dwellings or households, list of establishment in an Establishment Survey and list of farms in Agricultural Survey. Units in a frame must be identifiable and non-overlapping.

Sometimes a suitable frame may not be available and need to be constructed. In such a case, the procedure for construction should be clearly described. Even when a frame is available, it may be defective, incomplete, obsolete or full of duplications. The defective frame should be corrected for defects before it can be used for sample selection.

An incomplete frame is one in which some units that are supposed to be in the listing are missing. One way of solving the problem of incomplete frame is to redefine the population to be covered, provided the survey objective permits such a change.

Another method is to supplement the list from other existing frame. In case of duplication, any unit appearing more than once if known should have one of the duplicates retained and the rest deleted. Effort should be made to update an obsolete or out-of-date frame.

4. The information to be collected: The question of the kind of information to be collected should be considered at an early stage of planning the survey. Only data relevant to the purposes of the survey should be collected. If there are too many questions, the respondents begin to lose interest in answering them. On the other hand, it must be ensured that no important item is missing. A major consideration would be the practicability of obtaining the information sought.

5. Method of Data Collection Data may be collected through mail, questionnaire, personal interview, physical observation or vital registration depending on the type of data to be collected, unit of inquiry, the subject matter and the scale of the survey. The method of collecting information (whether by mail, or by interview or otherwise) has to be decided, bearing in mind the costs involved and level of accuracy desired.

For survey among educated people or survey of individual establishments mail questionnaire may be appropriate provided that addresses of such individuals are available.

Mail questionnaire is less expensive but has more non-response, especially in developing countries where postal system is very inefficient. In a survey where a large direct personal interview is recommended, it has less non-response. Data collected through personal interview are generally more reliable. Any combination of these methods could be used in any one survey.

5. Time reference and reference period: A decision has to be made concerning the time reference (period to which the results of the survey will relate) and the reference period (the period for which information is collected from sample units). For example, in one household survey the time reference was one year, but the reference period for most of the items was one week (each household was required to provide information for just one week). The choice of the reference period is important.

A shorter reference period may give more accurate data, but a larger sample is necessary with this method, and this means increased costs. A longer reference period may be cheaper, but the information collected may not be so accurate, owing to memory loss, etc.

Periods used in Survey:

i. **Survey Period:** This is the time period stipulated for the collection of information from the sample units, i.e., the period within which the survey is to be carried out. The survey period may last for 14 days, say July 3, 2011 to July 17, 2011.

ii. **Time reference:** This is the time period to which the results of the survey should refer. E.g. in an actual survey, the reference period may be a whole year.

iii. **Reporting Period/reference period**: This is the period for which the necessary information is collected for a sample unit at a time.

6. Organization of Fieldwork: The organization of fieldwork and supervision is very essential for a successful conduct of any survey. Much importance should be attached to the recruitment and training of good quality field staff. Field interviewers must be well groomed on the objectives of the survey and method of data collection.

They should also be well educated on the method of handling non-response. Trained supervisors should be used to monitor the field staff on a random basis during the course of the fieldwork in order to improve the quality of data collected. Supervision could be done through spot-checks, post survey checks and field scrutiny of the complete questionnaires.

7. Pre-test and Pilot Survey: Pre-test and pilot survey is a trial run for the main survey. It is very useful to try out questionnaires, field methods and to familiarize the interviewers on the job before the main survey. Pre-test and pilot survey has the following advantages:

a. Give insight into the population variability where this is not known. The population variability is used to determine the sample size needed to achieve the desired precision.
b. Access the adequacy of the questionnaire, which is the main function of pilot survey. In recoded questions, alternative answers may be discovered during pilot survey.
c. Ascertain the possible performance of the interviewers and areas of further training where necessary.
d. Give an idea of the non-response rate. Also, possible causes of non-response and effectiveness of the methods of treating non-response are known.
e. Help to ascertain the possible cost and duration of the main survey.

f. Help the researcher discover some of the objectives of the survey that are unrealizable. This may be as a result of respondents' unwillingness or inability to supply the required information.

8. Precision: It is necessary to specify the desired degree of precision beforehand for the various parameters to be estimated. The precision of an estimate can be improved upon by taking larger samples if cost permits, using efficient measuring instruments and good sample survey design.

9. Sampling Design: The choice of the optimum sampling design depends to a large extent on the cost of the survey, degree of precision and operational convenience. The general principle in the choice of any sampling design is to reduce cost for a given precision, or to reduce precision for a given cost of the survey while at the same time controlling the non-sampling errors.

10. Analysis of Survey Data: The first step in the analysis of survey data is to edit the completed questionnaires in order to delete and correct errors. To eliminate or minimize non-sampling errors, it is necessary to check thoroughly the transcription of data into summary forms, files, computational equipment and the summary tables.

Each stage of computation must be satisfactorily crosschecked. The objectives of the survey must be taken into consideration in the summarization of data into tables. Required estimates and their sampling errors should be presented. Also, cross-tabulation of data by geographical areas, demographic or socio-economic characteristics may be carried out.

11. Report Writing: The report of the results of a survey should be prepared under the following headings:

i. Objective of the survey: indicating the purpose of the survey.
ii. Scope: indicating the domains of study and the geographical areas covered by the survey.
iii. Subject coverage: a detailed description of the items of information collected should be given.

iv. Method of data collection: this should include a clear description of the method used in the collection of data; difficulties encountered and how they were surmounted

v. Survey, reference and reporting periods: these should be stated in the report.

vi. Sampling design and estimation procedure: the description should specify the sample size and how it was determined, sampling unit, the frame, methods of sample selection and estimation of parameters.

vii. Presentation of results: tables used in the presentation must bear title and column heading. diagrams, charts and graphs could be used to give a clearer picture of the survey results.

viii. Accuracy: the accuracy attained in the result should be stated together with the extent and nature of non-response and how it was treated.

ix. Sponsoring organization: this should be mentioned in the report.

In-Text Question

The following are periods used in survey except _____

A. Survey periods
B. Time Reference
C. Reporting Period
D. Dangling periods

In-Text Answer

D. Dangling periods

1.4 Activities that could involve sample survey

In its broadest sense the purpose of a sample survey is the collection of information to satisfy a definite need. The need to collect data arises in every conceivable sphere of human activity. For example:

Population. Most governments nowadays collect information regularly about the total population (number of persons); its distribution by area, sex, age, and other socio-economic characteristics; the rate of growth of the population; internal migration and so on.

Figure 1.1: population

These data help in determining the future needs for such items as food, clothing, shelter, education, recreational facilities. Broadly speaking, data on the nature and size of the population can be used for determining the demand for goods and services and the size and quality of the labour resources needed to produce these goods and services.

(ii) Labour: Since labour is a key resource in production, data are collected on the number of persons engaged in economic activity, the number of hours they work, and the average output per man-hour of work.

Data about the distribution of the labour force by branch of economic activity give a useful indication of the structure of production in the country.

Classifications of the economically active by occupation can be used to study the capabilities of the labour force from the point of view of development projects. Detailed information on the unemployed persons is used to find out what type of work they are looking for, how long they have been in search of work, what type of training they have had, and so on.

(iii) Agriculture: With rising populations, it is becoming more and more important to assess the agricultural resources of the country. The proportion of land under agriculture, areas under different crops, areas under pastures and forests, production of food- grains, fruits, etc.-- and the number and quality of livestock are some of the items of information essential to any planned programme of national development.

Figure 1.2: Agriculture

Data on the number and area of farm holdings by size and type of tenure can be used to determine the extent to which these factors may be contributing to agricultural productivity as well as to device remedies.

Summary for Study Session1

In study session 1, you have learnt that:

1. Sample is finite sequence of elements drawn from the population
2. **Sampling** is a scientific method of selecting and using a representative part (Sample) of a whole to seek the truth about the whole. Sampling is used extensively, consciously or unconsciously in everyday life to obtain the required information or to carry out a course of action.
3. **A sample survey** is defined as the collection and examination of data from a sample in order to make inferences about the whole. This is contrary to a census, which is a complete enumeration or survey of the whole units of enquiring.
4. The objectives of the survey should define the population the survey is intended to cover. But practical difficulties in handling certain segments of the population may point to their elimination from the scope of the survey.

5. Pre-test and pilot survey is a trial run for the main survey. It is very useful to try out questionnaires, field methods and to familiarize the interviewers on the job before the main survey.
6. Data about the distribution of the labour force by branch of economic activity give a useful indication of the structure of production in the country.

Self-Assessment Questions (SAQs) for Study Session 1

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 1.1 (Testing Learning outcomes 1.1)

Define sampling scheme

SAQ 1.2 (Testing Learning outcomes 1.2)

Mention the advantages of sample survey

SAQ 1.3 (Testing Learning outcomes 1.3)

Identify the periods used in survey

SAQ 1.4 (Testing Learning outcomes 1.4)

Explain the application of sample survey to agriculture

Notes on SAQ

SAQ 1.1

Sampling Scheme: This is a method of selecting a sample from a population, e.g. Simple random sampling, systematic sampling, etc.

SAQ 1.2

1. In destructive investigations, such as determination of the mean, Life-time of electric bulbs or the shelf life of some perishable items, it is better to use sample survey rather than complete examination.

2. Sample survey saves time, labour and cost, especially when these resources are limited, than a census where a large number of these resources is needed.

3. Data can be collected and analysed quickly from a sample than from the entire population.

4. There is a greater and more efficient supervision of field staff in a sample survey resulting in the collection of more reliable data.

5. Sample survey makes use of better qualified staff and specialized equipment.

6. It has greater subject coverage and less observational error than census.

SAQ 1.3

Periods used in Survey:

i. **Survey Period:** This is the time period stipulated for the collection of information from the sample units, i.e., the period within which the survey is to be carried out. The survey period may last for 14 days, say July 3, 2011 to July 17, 2011.

ii. **Time reference:** This is the time period to which the results of the survey should refer. E.g. in an actual survey, the reference period may be a whole year.

iii. **Reporting Period/reference period**: This is the period for which the necessary information is collected for a sample unit at a time.

SAQ1.4

With rising populations, it is becoming more and more important to assess the agricultural resources of the country. The proportion of land under agriculture, areas under different crops, areas under pastures and forests, production of food- grains, fruits, etc.-- and the number and quality of livestock are some of the items of information essential to any planned programme of national development.

Data on the number and area of farm holdings by size and type of tenure can be used to determine the extent to which these factors may be contributing to agricultural productivity as well as to device remedies.

References

Cochran, W.G, (1977); *Sampling Techniques* third edition, New York: John Wiley & Sons

Daroga Singh and Chaudhary F.S, (1986); *Theory and Analysis of Sample Survey Design*, New Delhi: Wiley Eastern Limited

Des Raj and Promod Chandhok (1998); *Sample Survey Theory*, New Delhi: Narosa Publishing House

Kish L. (1965); *Survey Sampling*, New York: John Wiley & Sons

Okafor F.C (2002); *Sampling Survey Theory with Applications*, Nsukka: Afro-Orbis Publishers

Mukhopadhyay P. (2005): *Theory and Methods of Survey Sampling*, New Delhi: Prentice-Hall of India Private Limited

Study Session 2: Probability and non-Probability sampling

Introduction

Survey research involve two methods. It involves two sampling methods. The two main kinds of sampling are probability (random) sampling and non-probability (non-random) sampling. In probability sampling, every unit of enquiry in the population has a known non-zero probability of being included in the sample.

This probability could be equal or unequal and in non- probability sampling, the probability of selecting a unit in the sample is not known and cannot be determined. In other words, a unit in the sample is not selected with any known probability. Hence, statistical inference cannot be made objectively about the population in non-random sampling.

Learning Outcomes for Study Session 2

At the end of this study, you should be able to:

2.1 Explain probability sampling

2.2 Explain non-probability sampling

2.1 Probability Sampling (Random)

Probability (random) sampling is a type of sampling and can be divided into two. They are:

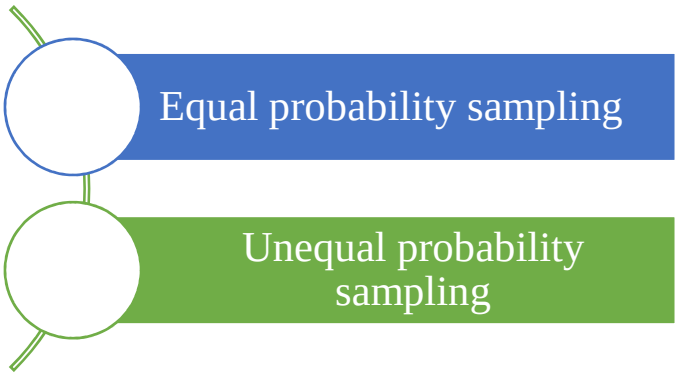


Figure 2.1: Probability Sampling

1. **In equal probability sampling**, every element in the population has the same non-zero probability of being included in the sample.
2. **Two probability sampling techniques used** for selecting a specified number of units from a population such that every unit has an equal chance of being in the sample are:

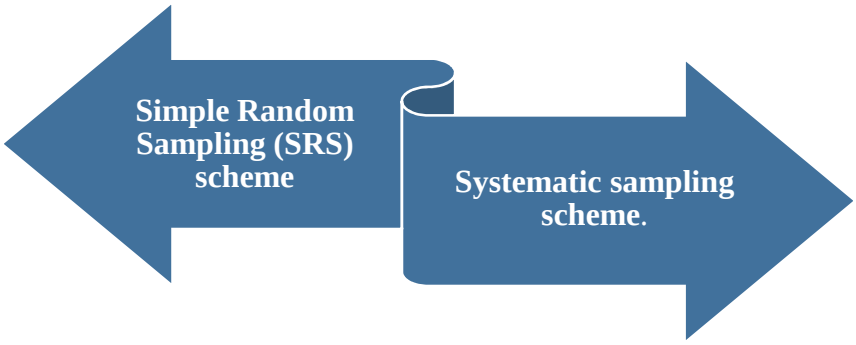


Figure 2.2: Two Probability sampling techniques

2.2.1 Simple Random Sampling (SRS)

Simple random sampling can be done with or without replacement.

Introduction

Simple random sampling is the simplest probability sampling procedure. In this method of selecting a sample of the population units, every sample of a fixed size is given an equal chance to be selected.

Every population unit is given an equal chance of appearing in the sample. Similarly, every pair of units has an equal chance of appearing in the sample. In general, every collection of units of a fixed size has an equal chance of being selected.

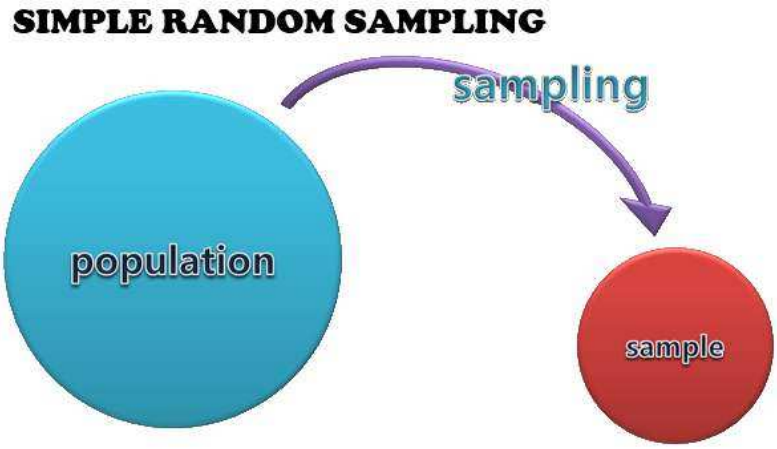


Figure 2.3: Simple random sampling

In the following sections, procedures for selecting a simple random sample and for using it to estimate the population means, totals, and variances of the characteristics of interest are presented. Methods for determining the sample size required for a survey are also examined.

Procedure

This fundamental method of sample selection may be described thus. From a population of N units, select one by giving equal probability to all units. This is best done with the help of random numbers.

Number the whole units in the population serially from 1 to N. Then select random numbers between 1 to N with the aid of table of random numbers, starting from the top of the column or columns, and depending on the number of digits that make up the population size N, to the bottom.

If the total random numbers required are not selected after getting to end of the column, use the next column or columns until the required numbers are obtained. The units where serial numbers correspond to the random numbers selected constitute the sample units.

Alternatively, the numbers 1 to N could be written out in pieces of paper or any suitable device. A number is drawn one after the other as in a lottery until the required number of sample units is obtained. This is called lottery method.

If after each draw the selected number is discarded before the next selection is made, the selection method is described as simple random sampling without replacement (SRSWor), generally referred to as simple random sampling (SRS).

On the other hand, if the selected number is replaced in the population before the next draw is made, it is described as simple random sampling with replacement (SRSWr). In SRSWr there is a possibility that a unit may appear more than once in the sample.

This is not the case with SRSWor where a unit appears once and only once in the sample. Thus, in SRSWr the sample size could exceed the population size. The sample obtained by SRSWr is called an “unrestricted or ordered sample” while that from SRSWor is called a “restricted or unordered sample”. Simple random sampling without replacement has a hyper-geometric probability distribution while SRSWr, follows a binomial distribution.

In-Text Question

Simple random sampling is the simplest probability sampling procedure. True or False

In-Text Question

True

2.2.2 Selection Probability

Let the sample size be n. The number of ways of selecting a distinct (different) unit number of N without replacement is ${}^{N}C_{n}$. Any one of these ${}^{N}C_{n}$ possible samples of size n has the same probability $\frac{1}{{}^{N}C_{n}}$ of being selected. The probability that any specified unit is included in the sample of size n is $\frac{n}{N}$, the sum of the probabilities of including the unit at the first draw, second draw, … n^{th} draw.

In SRSWr there are N^{n} possible samples of size n, each of which can be chosen with probability $\frac{1}{N^{n}}$ considering the order of appearance of the sample units. Since successive draws are independent of one another, the probability of selecting a specified unit u_i at any of the draws is $\frac{1}{N}$. Hence the probability of any specified unit appearing in the ordered (unrestricted) sample of size n is also obtained by summing the probability $\frac{1}{N}$ up to n times, which gives $\frac{n}{N}$.

2.2.3 Properties of the Estimates (SRS)

The precision of any estimate made from a sample depends both on the method by which the estimate is calculated from the sample data and on the plan of sampling.

1. **Unbiasedness:** A method of estimation is unbiased if the average value of the estimate, taken over all possible samples of given size n, is exactly equal to the true population value.

To investigate whether $\bar{y}$ is unbiased with SRS; we calculate the value of $\bar{y}$ for all ${}^{N}C_{n}$ samples and find the average of the estimates.

2. **Consistency**: A method of estimation is called consistent if the estimate becomes exactly equal to the population value when n = N i.e., when the sample consists of the whole population. For SRS, $\bar{y}$ and $N\bar{y}$ are consistent estimates of the population mean and total respectively. An estimator is consistent if the probability that it is in error by more than any given amount tends to zero as the sample becomes large.

2.2.4 Estimation of Population Mean

Let Y_i be the value of the character y associated to each of the population units U_i. In order to estimate the population mean, we select a simple random sample of size n. Information on the character y is then obtained from each unit in the sample. Let $y_i (i = 1, 2, ...n)$ be the value of the character obtained from the ith sample unit. Then the sample mean

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \text{.....(2.0)}$$

is an estimator of the population mean μ. The sample mean $\bar{y}$ is an unbiased estimator of the population mean $\bar{Y}$

Proof

To show that $\bar{y}$ is unbiased we first of all recall that the expected value of a sum is the sum of the expected values and that the probability of selecting the i^{th} unit at any draw is $\frac{1}{N}$. Hence,

$$E(\bar{y}) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{N}\sum_{i=}^{N} Y_i\right) = \bar{Y}$$

.

Its variance is given by

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

where $S^2 = \dfrac{\sum (y_i - \bar{y})^2}{n-1}$

$f = \dfrac{n}{N}$ = sampling fraction (population correction factor).

Its inverse $\dfrac{\mathrm{N}}{\mathrm{n}}$ is the raising or inflation factor.

} *SRSWor*.....(2.1)

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

Proof

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

Multiplying both sides by n^2

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

$$= E\left\{\sum_{i=1}^{n} (y_i - \bar{Y})^2 + \sum_{i=1}^{n}\sum_{j\neq i}^{n} (y_i - \bar{Y})(y_j - \bar{Y})\right\}$$

In the case of SRSWr, the second term in the chain brackets on the right hand side vanishes as it is the covariance term. The selections at the i^{th} and j^{th} draws are independent.

Therefore,

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

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

Dividing through by n^2

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

For SRSWor, the selection of units is not independent and the covariance term does not varnish. Therefore,

$$n^2V(\bar{y}) = n\sigma^2 + E\left\{\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left(y_i - \bar{Y}\right)\left(y_j - \bar{Y}\right)\right\}$$

Since the probability of selecting the pair $(U_i, U_j), i \neq j$, together is $\frac{1}{N}\times\frac{1}{N-1}$ and the number of pairs of units in a sample of size n is $n(n-1)$. Also the probability that the pair (U_i, U_j) is together in the sample is $\frac{n(n-1)}{N(N-1)}$. Hence

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

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

The term in the chain brackets varnishes so that we now have

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

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

Dividing through by n^2 the final result becomes

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

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

2.2.5 Estimation of Population Total

The estimator of the population total Y is given by

$\hat{Y} = N\bar{y}$, with variance of $\hat{Y}$

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

or

$$V\left(\bar{Y}\right) = N^2\left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} = N^2\left(\frac{1-f}{n}\right)S^2 \text{ for SRSWOR}$$

2.2.6 Estimation of the Standard Error from a Sample

The standard errors of the estimated population mean and total area are used primarily for three purposes:

- To compare the precision obtained by SRS with that given by other methods of sampling;
- To estimate the size of the sample needed in a survey that is being planned;
- To estimate the precision actually attained in a survey that has been completed.

The formula involves S^2, the population variance. In practice, this is not usually known but can be estimated from the sample data.

Corollary 1

The standard error of $\bar{y}$ is

$$\sigma_{\bar{y}} = \frac{S}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}} = \frac{S}{\sqrt{n}}\sqrt{1-f}$$

Corollary 2

The variance of $\hat{y} = N\bar{y}$, as an estimate of the population total Y, is

$$\mathrm{V}\left(\bar{Y}\right) = E\left(\bar{Y}-Y\right)^2 = \frac{N^2S^2}{n}\frac{\left(N-n\right)}{N} = \frac{N^2S^2}{n}\left(1-f\right)$$

Corollary 3

The standard error of $\hat{y}$ is

$$\sigma_{\hat{y}} = \frac{NS}{\sqrt{n}}\sqrt{\frac{\left(N-n\right)}{N}} = \frac{NS}{\sqrt{n}}\sqrt{1-f}$$

In-Text Question

The standard errors of the estimated population mean and total area are used primarily for these purposes except:

A. To compare the precision obtained by SRS with that given by other methods of sampling;
B. To estimate the size of the sample needed in a survey that is being planned
C. To estimate the size of mean data
D. To estimate the precision actually attained in a survey that has been completed.

In-Text Question

C. To estimate the size of mean data

2.2.7 The Finite Population Correction

For a random sample of size n from an infinite population, it is well known that the variance of the mean is $\frac{\sigma^2}{n}$. The only change in this result when the population is finite is the introduction of the factor $\frac{N-n}{N}$. The factors $\frac{N-n}{N}$ for variance and $\sqrt{\frac{N-n}{N}}$ for the standard error are called the finite population correction (fpc).

Provided that the sampling fraction $\frac{n}{N}$ remains low, these factors are close to unity, and the size of the population as such has no direct effect on the standard error of the sample mean.
In practice, fpc can be ignored whenever the sampling fraction does not exceed 5% and for many purposes even if it is as high as 10%. The effect of ignoring the correction is to overestimate the stranded error of the estimate $\bar{y}$.

Example:

Consider the weights at birth y, of a population of four babies delivered in one day in a maternity clinic.

U_i	1	2	3	4
Y_i	3.9kg	3.6kg	3.4kg	3.7kg

(a) Select all possible samples of size n = 2

(i) With replacement

(ii) Without replacement

(b) Obtain the population mean

(c) Obtain the sample mean

(d) Obtain the sample variance/population variance

Solution

ai) All samples of size 2 in SRSWr

Samples	**Sample units**	**Probability**	**Sample mean**
1	(3.9,3.9)	1/16	3.9
2	(3.9,3.6)	1/16	3.75
3	(3.9,3.4)	1/16	3.65
4	(3.9,3.7)	1/16	3.8
5	(3.6,3.9)	1/16	3.75
6	(3.6,3.6)	1/16	3.6
7	(3.6,3.4)	1/16	3.5
8	(3.6,3.7)	1/16	3.65
9	(3.4,3.9)	1/16	3.65
10	(3.4,3.6)	1/16	3.5

11	(3.4,3.4)	1/16	3.4
12	(3.4,3.7)	1/16	3.55
13	(3.7,3.9)	1/16	3.8
14	(3.7,3.6)	1/16	3.65
15	(3.7,3.4)	1/16	3.55
16	(3.7,3.7)	1/16	3.7
Mean			**3.65**

aii) All samples of size 2 in SRSWor

Samples	**Sample units**	**Probability**	**Sample mean**	$\frac{(N-n)}{N}\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n(n-1)}$
1	(3.9,3.6)	1/6	3.75	0.01125
2	(3.9,3.4)	1/6	3.65	0.03125
3	(3.9,3.7)	1/6	3.8	0.005
4	(3.6,3.4)	1/6	3.5	0.005
5	(3.6,3.7)	1/6	3.65	0.00125
6	(3.4,3.7)	1/6	3.55	0.00125
Mean			**3.65**	**0.01083**

b) The population mean is obtained using the weights of all the four babies.

$$\bar{Y} = \frac{3.9+3.6+3.4+3.7}{4} = \frac{14.6}{4} = 3.65$$

$$\sigma^2 = 0.0325$$

c) From (ai) and (aii) we discover that the mean of the sample means, both for sampling with replacement and without replacement, are equal to the population mean. Therefore, we say that the sample mean is an unbiased estimator of the population mean.

d) i. The mean sample variance for SRSWr is given by

$$\sigma_{\bar{y}}^2 = V(\bar{y}) = \frac{\sigma^2}{n} = \frac{0.0325}{2} = 0.01625$$

ii. The mean sample variance for SRSWor is given by

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

Where

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

See column 5 of table (aii).

2.2 Non-Random Sampling (non probability)

Non probability sampling can be categorized into the following:

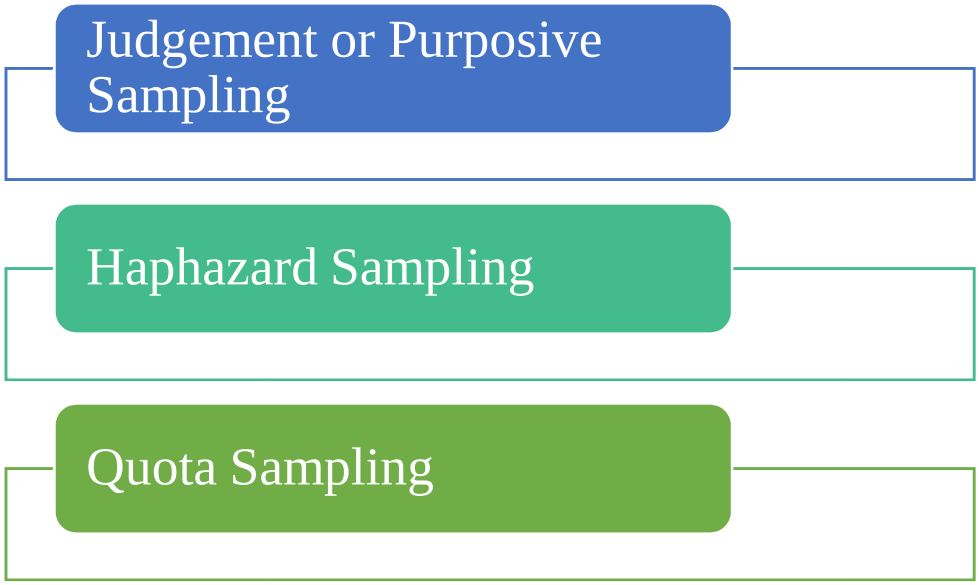


Figure 2.4: Non Probability Sampling categories

(i) Judgment or Purposive Sampling

In this type of non-random sampling experts pick units they judge to be representative of the population. E.g. a typical town may be picked to represent an urban population or a village chosen to represent a rural population.

Since experts differ in their judgments, different experts may choose different units, which to them may appear to be representative of the population. Confidence is, therefore, not placed on the results from judgment sampling since there is no objective method for preferring one judgment to another.

There is always a danger of making systematic errors of judgment. In judgment sampling, a sampler may decide to choose villages closest to him or those having a particular characteristic. Finally, accurate measure of sample variability cannot be obtained with judgment sampling.

(ii) **Haphazard Sampling:** In haphazard sampling, units are taken into the sample as they come along or make themselves available. In a hospital survey on the characteristics of patients suffering from a particular disease, the patients are taken into the sample in the order they report to the hospital and volunteer to be part of the study. This sampling method lacks representativeness of the population to be studied.

(iii) **Quota Sampling:** This is in a way similar to purposive sampling. In quota sampling, specified number of sample units in each specified group, based on say, gender, age, social class, etc. is assigned to an enumerator. The enumerator then selects the required representative number of individuals to be in the sample within each quota (group).

Quota sampling is used extensively in opinion surveys and in market research. The disadvantage of quota sampling is that it may not be possible for the interviewer to fill the quota with representative units. Another disadvantage is that sampling variability cannot be calculated with quota sampling.

The advantage of quota sampling is that it is easy to apply and avoids the problem of not-at-homes and call-backs. It is useful in obtaining quickly the people's reaction to current issues. Quota sampling can be used even when there is no suitable list of population units.

In-Text Question

The disadvantage of quota sampling is that it may not be possible for the interviewer to fill the quota with representative units. True or False

In-Text Answer

True

Summary for Study Session 2

In study session 2, you have learnt that:

1. **In equal probability sampling**, every element in the population has the same non-zero probability of being included in the sample.
2. **Two probability sampling techniques used** for selecting a specified number of units from a population such that every unit has an equal chance of being in the sample are:
3. Simple random sampling is the simplest probability sampling procedure. In this method of selecting a sample of the population units, every sample of a fixed size is given an equal chance to be selected.
4. In haphazard sampling, units are taken into the sample as they come along or make themselves available. In a hospital survey on the characteristics of patients suffering from a particular disease, the patients are taken into the sample in the order they report to the hospital and volunteer to be part of the study. This sampling method lacks representativeness of the population to be studied.
5. Quota sampling is used extensively in opinion surveys and in market research. The disadvantage of quota sampling is that it may not be possible for the interviewer to fill the quota with representative units. Another disadvantage is that sampling variability cannot be calculated with quota sampling.

Self-Assessment Questions (SAQs) for Study Session 2

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 2.1

Define in equal probability sampling

SAQ 2.2

Explain quota sampling

Notes on SAQ

SAQ 2.1

In equal probability sampling, every element in the population has the same non-zero probability of being included in the sample.

SAQ 2.2

Quota Sampling: This is in a way similar to purposive sampling. In quota sampling, specified number of sample units in each specified group, based on say, gender, age, social class, etc. is assigned to an enumerator.

References

Cochran, W.G, (1977); *Sampling Techniques* third edition, New York: John Wiley & Sons

Daroga Singh and Chaudhary F.S, (1986); *Theory and Analysis of Sample Survey Design*, New Delhi: Wiley Eastern Limited

Des Raj and Promod Chandhok (1998); *Sample Survey Theory*, New Delhi: Narosa Publishing House

Kish L. (1965); *Survey Sampling*, New York: John Wiley & Sons

Okafor F.C (2002); *Sampling Survey Theory with Applications*, Nsukka: Afro-Orbis Publishers

Mukhopadhyay P. (2005): *Theory and Methods of Survey Sampling*, New Delhi: Prentice-Hall of India Private Limited

Study Session 3: Estimation of Population Proportion

Introduction

Population can be described as the part of a population with a definite characteristics conveyed as a fraction, decimal or percentage of the whole population. In order to estimate the proportions of some attribute within a population, it is necessary that you depend on the proportions observed within a sample of the population.

The sample should be a random sample. Size of the sample is important. As the random sample increases, there is increase believe that observed sample proportion will be close to the actual population proportion. In this study session, you will learn about estimation of population proportion and the confidence limit in the estimation of population proportion.

Learning outcomes for study session 3

At the end of this study, you should be able to:

3.1 Explain the estimation of population proportion

3.2 Explain the confidence limit in the estimation of population proportion

3.1 Estimation of population proportion

Estimating population proportions can viewed as the particular case of estimating the population mean. Suppose that the population of N units can be classified into mutually exclusive classes or domains, or that the units possess defined attributes. It may, apart from estimating the population mean or total, be of interest to estimate the population proportion or number of units belonging to a given class.

For example, in an establishment, survey interest may be in knowing the proportion or the number of establishments with less than a given number of employees. Also, in a rural agricultural Survey we may want to know the proportion of rural farmers who use fertilizer or who belong to Farmers' Cooperative Society.

Let

N_c is the number of units in the population of size N belonging to a given class or domain D.

$P = \frac{N_c}{N}$ is the proportion of units in the population belonging to D.

Consider a simple random sample of size n selected from N units with or without replacement. The estimator of the population proportion P is given by

$P = \frac{n_c}{n}$, called the sample proportion.

n_c is the number of units in the sample belonging to class D.

The sampling variance of P is

$V(P)$ $\frac{\text{PQ}}{\text{n}}$ for SRSWr

OR

$\text{V(P)} = \frac{\text{N-n}}{\text{N-1}} \frac{\text{PQ}}{\text{n}} = \frac{1-f}{n} \frac{\text{PQ}}{\text{N-1}}$ for SRSWor

Where Q = 1- P

$\hat{V}(P) = \frac{1-f}{n-1} pq$

In-Text Question

Estimating population proportions means estimating the population mean. True or False

In-Text Answer

True.

3.1.1 Estimation of N_c

The estimator of the number of units N_c in the population belonging to a given class is

$$\hat{N}_c = Np$$

with variance

$$V\left(\hat{N}_c\right) = N^2 \frac{PQ}{n} \quad \text{for SRSWr}$$

or

$$\mathrm{V}\left(\hat{\mathrm{N}}_{\mathrm{c}}\right) = N^2 \left(\frac{\text{1-f}}{\text{n}}\right) \frac{\text{NPQ}}{N-1} \quad \text{for SRSWor}$$

Example: From a hospital admission record, a sample of 120 in-patients was selected by SRSWor from 1556 in-patients admitted in the hospital in one particular year. Information on the sex of each patient and the number of days a patient was hospitalized before being discharged was obtained. Suppose that the mean length of stay and the proportion of females admitted in the hospital are of interest to the investigator.

From the sample observation, the total length of stay (in days) of the 120 in-patients is;

$$\sum_{i=1} y_i = 1567; \quad \sum_{i=1} y_i^2 = 44121$$

The number of females in the sample is 92.

Solution

Mean length of stay:

$$\bar{y} = \frac{1567}{120} = 13.058 \,\text{days}$$

Estimate of the Sampling Variance of $\bar{y}$

Since the population variance S^2 is unknown, its sample estimate is

$$\hat{s}^2 = \frac{44121 - 120\ (13.058)^2}{119} = 198.820$$

Using

$$\hat{V}\left(\bar{y}\right) = \left(1 - f\right)\frac{\hat{s}^2}{n}$$

$$= \left(1 - \frac{120}{1556}\right) \text{X } \frac{198.820}{120} = 1.529$$

The estimated standard error (S.E) is

$$Se(\bar{y}) = \sqrt{\hat{V}(\bar{y})} = \sqrt{1.529} = 1.237$$

The 95% C.I. for the mean, assuming normal distribution, is

$$13.058 \pm 1.96 \times 1.237 = 13.058 \pm 2.425$$

Estimate of the Proportion of Females:

$$P = \frac{n_c}{n} = \frac{92}{120} = 0.767$$

The estimate of the sampling variance of P is obtained by using the relation

$$\hat{V}(P) = \frac{1-f}{n-1} pq$$
$$= 1 - \frac{120}{1556} \text{ x } \frac{0.767 \text{ x } 0.233}{119} = 0.0014$$

The estimate of the number of females admitted in the hospital is

$$\hat{N}_c = Np$$
$$= 1556 \text{ x } 0.767 = 1193 \text{ females}$$

Its variance estimate is given as

$$\hat{V}(\hat{N}_c) = N^2 \hat{V}(p)$$
$$= (1556)^2 \ (0.0014) = 3389.56$$

3.2 Confidence Limit

It is usually assumed that the estimates $\bar{y}$ and $\hat{Y}$ are normally distributed about the corresponding population values. If the assumption holds, lower and upper confidence limits for the population mean and total are as follows:

Mean:

$$\hat{\bar{Y}}_L = \bar{y} - \frac{Z_\alpha s}{\sqrt{n}} \sqrt{1-f}, \hat{\bar{Y}}_U = \bar{y} + \frac{Z_\alpha s}{\sqrt{n}} \sqrt{1-f}$$

Total:

$$\hat{Y} = N\overline{y} - \frac{Z_\alpha NS}{\sqrt{n}}\sqrt{1-f}\,, \hat{Y} = N\overline{y} + \frac{Z_\alpha NS}{\sqrt{n}}\sqrt{1-f}$$

The symbol Z_α is the value of the normal deviate corresponding to the desired confidence probability. The most common values are

Confidence probability (%)	**50**	**80**	**90**	**95**	**99**
Value	0.67	1.28	1.64	1.96	2.58

If the sample size is less than 50, the percentage points may be taken from student's $t-table$ with $(n-1)$ degrees of freedom. The t-distribution holds exactly only if the observations y_i are normally distributed and N is infinite.

Example: Signatures to a petition were collected on 676 sheets. Each sheet had enough space for 42 signatures, but on many sheets a smaller number of signatures had been collected. The numbers of signatures per sheet were counted on a random sample of 50 sheets (about 7% sample), with the following results.

y_i	42	41	36	32	29	27	23	19	16	15	14	11	*10*
f_i	23	4	1	1	1	2	1	1	2	2	1	1	*1*
y_i	9	7	6	5	4	3							
f_i	*1*	*1*	*3*	*2*	*1*	*1*							

y_i Number of signature

f_i Frequency

Estimate the total number of signatures to the petition and the 80% confidence limits.

Solution

We believe that the samples are normally distributed.

We find:

$$n = \sum f_i = 50, \quad \sum f_i y_i = 1471, \quad \sum f_i y_i^2 = 54,497, \quad N = 676$$

Hence the estimated total number of signatures is

$$\hat{Y} = N\ \bar{y} = \frac{(676)(1471)}{50} = 19{,}888$$

For the sample variance $\hat{s}^2$ we have

$$S^2 = \frac{1}{n-1}\left[\sum f_i\ (y_i - \bar{y})^2\right] = \frac{1}{n-1}\left[\sum f_i y_i^2 - \frac{(\sum f_i y_i)^2}{\sum f_i}\right]$$

$$= \frac{1}{49}\left[54{,}497 - \frac{(1471)^2}{50}\right] = 229.0$$

The 80% confidence limits are

$$19{,}888 \pm \frac{Z_{0.2} NS}{\sqrt{n}}\sqrt{1-f} = \frac{19{,}889 \pm (1.28)\ (15.13)\ (676)}{\sqrt{50}}$$

$$= (18{,}107, \quad 21{,}669)$$

Summary for study sessionn3

In this study session, you have learnt that:

1. **Estimating population proportion**s can viewed as the particular case of estimating the population mean.

2. It is usually assumed that the estimates $\bar{y}$ and $\hat{Y}$ are normally distributed about the corresponding population values.
3. The estimator of the number of units N_c in the population belonging to a given class is

$$\hat{N}_c = Np$$

4. Lower and upper confidence limits for the population mean and total are as follows:
    a. Mean:

$$\hat{\bar{Y}}_L = \bar{y} - \frac{Z_\alpha s}{\sqrt{n}}\sqrt{1-f}\,, \hat{\bar{Y}}_U = \bar{y} + \frac{Z_\alpha s}{\sqrt{n}}\sqrt{1-f}$$

Self-Assessment Questions (SAQs) for Study Session 3

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 3.1

Explain the estimation of population proportion

Notes on SAQ

SAQ 3.1

Estimating population proportions can viewed as the particular case of estimating the population mean. Suppose that the population of N units can be classified into mutually exclusive classes or domains, or that the units possess defined attributes. It may, apart from estimating the population mean or total, be of interest to estimate the population proportion or number of units belonging to a given class.

References

Cochran, W.G, (1977); *Sampling Techniques* third edition, New York: John Wiley & Sons

Daroga Singh and Chaudhary F.S, (1986); Theory and Analysis of Sample Survey Design, New Delhi: Wiley Eastern Limited

Des Raj and Promod Chandhok (1998); *Sample Survey Theory*, New Delhi: Narosa Publishing House

Kish L. (1965); *Survey Sampling*, New York: John Wiley & Sons

Okafor F.C (2002); *Sampling Survey Theory with Applications*, Nsukka: Afro-Orbis Publishers

Mukhopadhyay P. (2005): *Theory and Methods of Survey Sampling*, New Delhi: Prentice-Hall of India Private Limited

Study Session 4: Systematic Sampling

Introduction

In the previous study session, you learnt about estimation of population proportion which is a case of determining the population mean. In this study session however, you will learn about systematic sampling which is a type of probability sampling method in which sample members from a bigger population are picked according to a random starting point and a fixed periodic interval.

You will also learn about the estimation in systematic sampling and the determination of sample size.

Learning Outcomes for study session 4

In study session 4, you should be able to:

4.1 Define systematic sampling
4.2 Explain the estimation of systematic sampling
4.3 Explain the determination of sample size

4.1 Systematic Sampling

Systematic sampling is a type of probability sampling method in which sample members from a bigger population are picked according to a random starting point and a fixed periodic interval.

Systematic sampling is a more convenient method of sample selection when the units are serially numbered from 1 to N. Suppose $N = nk$ where n the sample size is desired and k is an integer.

A number is taken at random from the numbers 1 to k (using a table of random numbers). Suppose the random number is i, then the sample contains n units with serial numbers $i, i+k, i+2k, \cdots, i+(n-1)k$.

Thus, the sample consists of the first unit selected at random and every k^{th} unit thereafter. This procedure of sample selection is called **systematic sampling** and the sample is called **systematic sample**. k is known as the sampling interval and $\frac{1}{k} = \frac{n}{N}$ is the sampling fraction.

The convenience of selection lies in the fact that the selection of the first member of the sample determines the entire sample automatically.

You need to note the following about this procedure.

1. The probability of selecting one group or sample or cluster of units from the k samples or groups or clusters is $\frac{1}{k}$.

2. Each unit U_i in the population belongs to one and only one group or sample or cluster and the probability of selecting a cluster, as earlier mentioned, is $\frac{1}{k}$. This shows that systematic sampling is a probability sampling procedure.

k -Systematic Possible clusters

Cluster	***Composition of cluster***
1	$1, \quad k+1, \quad 2k+1, \quad \cdots \quad (n-1)k+1$
…	
i	$i, \quad k+i, \quad 2k+i, \quad \cdots \quad (n-1)k+i$
…	
k	$k, \quad 2k \quad 3k, \quad \cdots \quad nk$

In case $N \neq nk$ some of the clusters or samples will contain n units while others will contain $n+1$units; i.e. sample sizes will not be equal. But the probability that a given unit is selected into the sample (of size $n+1$ or n) will still be $\frac{1}{k}$, since one sample is selected at random from the k possible samples.

In-Text Question

This procedure of sample selection is called systematic sampling. True or False

In-Text Answer

True

4.2 Estimation in Systematic Sampling

In systematic sampling with interval k, an unbiased estimator of the mean, $\bar{y}_{sy}$ is the same as the mean of the i^{th} systematic sample. This is given as

$$\bar{y}_{sy} = \bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

where

y_{ij} = value of the jth unit of the ith systematic sample

Thus, if $N = nk$, $\bar{y}_{sy}$ is unbiased for the population mean. It also reduces to the sample mean and can be written as:

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

where

y_i is the value of the ith unit of any systematic sample.

The variance of $\bar{y}_{sy}$ is the difference between the total variance σ^2 and within sample variance σ_w^2.

$$V(\bar{y}_{sy}) = \sigma^2 - \sigma_w^2$$

where

$$\sigma^2 = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$$

and

$$\sigma_w^2 = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2 = \frac{1}{k}\sum_{i=1}^{k}\frac{n-1}{n}S_{wi}^2 = \frac{n-1}{n}S_w^2$$

$$= \frac{k(n-1)}{N}S_w^2$$

By definition

$$S_{wi}^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2$$

The variance of $\bar{y}_{sy}$ can be expressed in the form

$$V(\bar{y}_{sy}) = \frac{N-1}{N}S^2 - \frac{k(n-1)}{N}S_w^2$$

where

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

Note that the variance of a systematic sample could be decreased if within sample variance S_w^2 is increased since the population variance, S^2 is constant for any given population. You can achieve this by ensuring that each systematic sample is internally heterogeneous. That is, units within each systematic sample vary as much as possible.

The relative efficiency of systematic sampling with respect to SRSW or will be greater or equal to unity according as $\rho_c \leq \frac{-1}{(N-1)}$. This implies that there must be a negative correlation between units within each systematic sample.

In other words, the units must be heterogeneous. If $\rho_c < \frac{-1}{(N-1)}$, systematic sampling will be more precise than SRSW or. When $\rho_c = \frac{-1}{(N-1)}$, systematic sampling will be as precise as SRSW or. This will be the case when units are in random order.

The performance of systematic sampling depends largely on the arrangement of the population units.

4.3 Determination of Sample Size

In the sample survey, the question is often, what is the adequate number of units needed in the sample from which the desired information is to be obtained? Too large a sample may result in the waste of resources, and too small a sample may render the survey result less precise and useless. Sample size depends on a number of factors, namely:

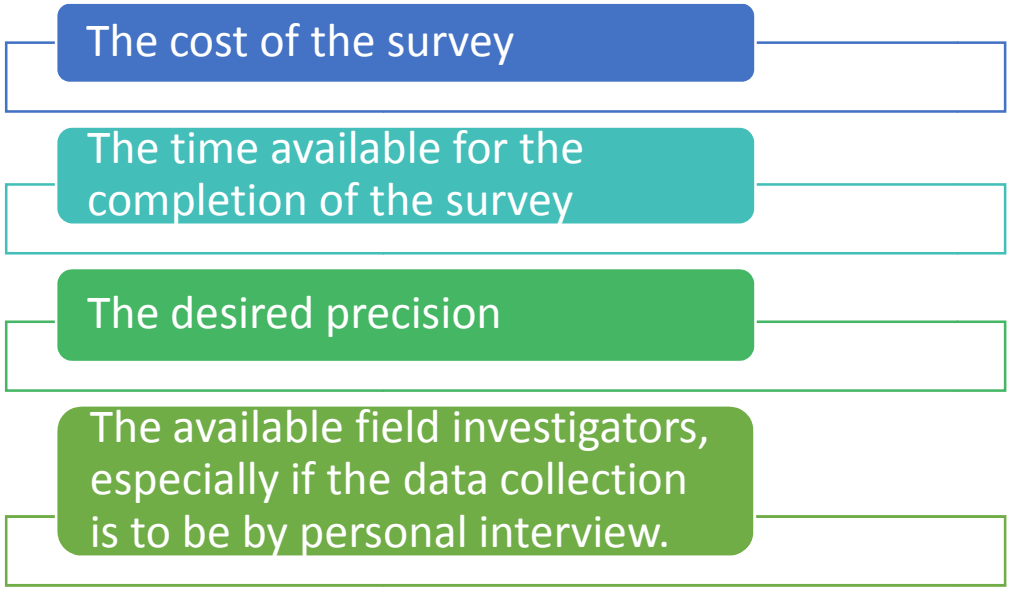


Figure 4.1: Factors that determine sample size

Thus, if enough fund and personnel are available for the survey, a restriction may be placed on the sample size by the time required to complete the survey. On the other hand if enough time, fund and personnel are available, we may be constrained in the choice of the sample size by the precision desired for the estimates. Sample size is, however, never a function of only one of these factors but interplay of all of them.

In-Text Question

Sample size depends on the following except _____

A. The cost of the survey

B. The time available for the completion of the survey

C. Cost of production

D. The desired precision

In-Text Answer

C. Cost of production.

Summary for Study Session 4

In study session 4, you have learnt that:

1. **Systematic sampling** is a type of probability sampling method in which sample members from a bigger population are picked according to a random starting point and a fixed periodic interval.
2. In systematic sampling with interval k, an unbiased estimator of the mean, $\bar{y}_{sy}$ is the same as the mean of the i^{th} systematic sample. This is given as

   $$\bar{y}_{sy} = \bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

   where
3. y_{ij} = value of the j^{th} unit of the i^{th} systematic sample
4. In the sample survey, the question is often, what is the adequate number of units needed in the sample from which the desired information is to be obtained?
5. . Sample size depends on a number of factors, namely:

a) The cost of the survey

b) The time available for the completion of the survey

c) The desired precision

d) The available field investigators, especially if the data collection is to be by personal interview.

Self-Assessment Questions (SAQs) for Study Session 4

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 4.1

Define systematic sampling

SAQ 4.2

Outline the factors that determine sample size

Notes on SAQ

SAQ 4.1

Systematic sampling is a type of probability sampling method in which sample members from a bigger population are picked according to a random starting point and a fixed periodic interval.

SAQ 4.2

- The cost of the survey
- The time available for the completion of the survey
- The desired precision
- The available field investigators, especially if the data collection is to be by personal interview.

References

Cochran, W.G, (1977); *Sampling Techniques* third edition, New York: John Wiley & Sons

Daroga Singh and Chaudhary F.S, (1986); *Theory and Analysis of Sample Survey Design*, New Delhi: Wiley Eastern Limited

Des Raj and Promod Chandhok (1998); *Sample Survey Theory*, New Delhi: Narosa Publishing House

Kish L. (1965); *Survey Sampling*, New York: John Wiley & Sons

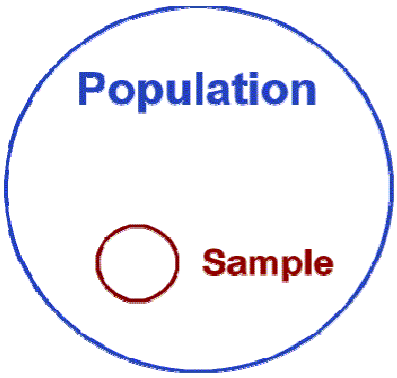
Okafor F.C (2002); *Sampling Survey Theory with Applications*, Nsukka: Afro-Orbis Publishers

Mukhopadhyay P. (2005): *Theory and Methods of Survey Sampling*, New Delhi: Prentice-Hall of India Private Limited

Study Session 5: Stratified Random Sampling

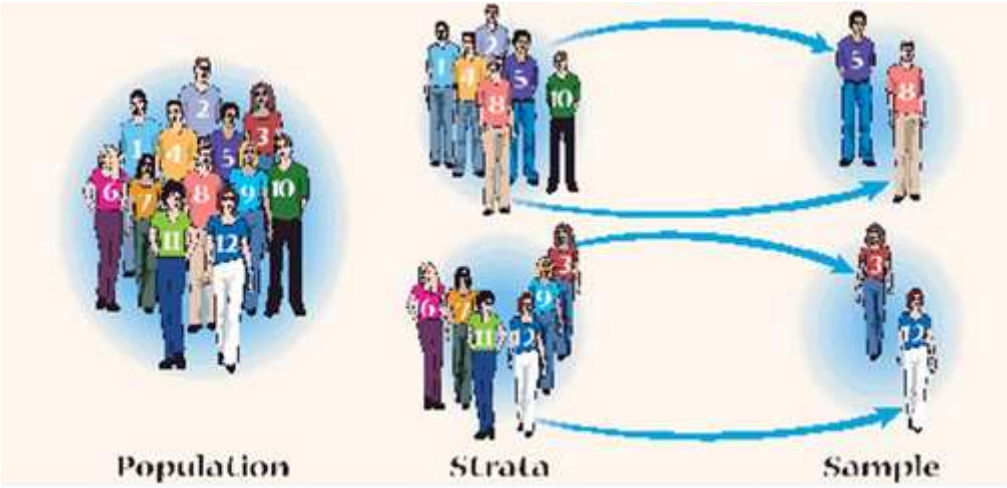
Introduction

In simple random sampling the variance of the estimate (say, of population mean $\bar{Y}$) depends, apart from the sample size, on the variability of the character y in the population. If the population is very heterogeneous and considerations of cost limit the size of the sample, it may be found impossible to get a sufficiently precise estimate by taking a simple random sample from the entire population.



And populations encountered in real life are generally very heterogeneous. For example, in surveys of manufacturing establishments, it can be found that some establishments are very large, i.e., they employ 1,000 or more persons, but there are many others which have only two or three persons on their pay rolls.

Any estimate made from direct random sample taken from the totality of such establishments would be subject to exceedingly large sampling fluctuations. But suppose it is possible to divide this population into parts (or strata) on the basis of, say, employment, thereby separating the very large ones, the medium-sized ones, and the smaller ones.



Source

If a random sample of establishments is now taken from each stratum, it should be possible to make a better estimate of the strata averages, which in turn should help in producing a better estimate of the population average. This is the basic consideration involved in the use of stratification for improving the precision of estimation.

Learning outcomes for Study Session 5

At the end of this study, you should be able to:

5.1 Explain stratified random sampling
5.2 Identify the reasons for stratification
5.3 Discuss estimation in stratified sampling
5.4 Explain the estimation of population mean
5.5 Explain the allocation of sample size to strata.

5.1 Stratified Random Sampling

Stratification is a method of using auxiliary information to increase the precision of the estimate of population characteristic. Suppose our interest is in estimating the total yield of a particular crop in a given geographical region, it may be advisable to group the list of farms according to sizes or any measure of size before selecting the farms.

This can be done such that small farms are in a group of their own, moderately sized farms in another group, while large farms are in a different group. If random samples of farms are now

drawn independently from each group, a more precise estimate of the total yield could be obtained without necessarily increasing the total sample size than when a random sample of the same sizes is selected directly from the entire population of farms in the region.

Box 5.1 Definition of stratified random sampling

Stratification is a method of using auxiliary information to increase the precision of the estimate of population characteristic.

If the farms are heterogeneous in size, the yield obtained from each farm will vary greatly according to the size of the farm. Consequently the population variability of the farm yields will be very large. If, however, the population of farms is divided into sub-populations called strata in such a way that farms within a stratum are homogeneous, then the stratum variability is reduced.

This requires independently sampling each stratum and, combining the stratum estimates to obtain the total precise estimate.

This procedure of drawing independent samples after grouping the whole units in the population into homogeneous distinct strata is called “stratified sampling”. When a simple random sampling is used to select sample in each stratum, the procedure is called “stratified random sampling”.

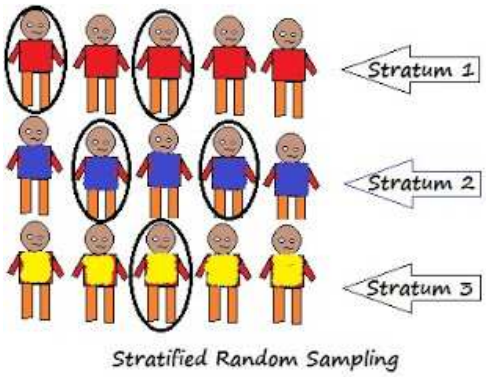


Figure 5.1: Stratified Random Sampling

If the main purpose of stratification is to achieve higher precision, a number of questions arise for which answers must be found. How should the strata be made and how many of

them should be made? How should the total sample be allocated to the strata? How should data be analysed (estimates be made and their variances calculated) from a stratified design?

In-Text Question

When a simple random sampling is used to select sample in each stratum, the procedure is called _____

A. Stratified random sampling
B. Random Sampling
C. Probability sampling
D. Sampling by the books

In-Text Answer

A. Stratified random sampling

5.2 Reasons for Stratification

Apart from increase in precision there are other reasons for stratification. Because some household characteristics differ in rural and urban areas, households may be classified into rural and urban, and different sampling scheme used for the selection of households in each stratum. Similarly, people living in institutions like hotels, barracks, prisons and hostels could be grouped in one stratum and those living in ordinary homes in another stratum.

Strata may be formed because the sub-populations in which the entire population is sub-divided could each be treated as a population in its own right and designated as a domain of study. For example, in a National Agricultural Sample Survey, (NASS), estimates may be needed for the whole country as well as for each Local Government Area (LGA).

In this case, the LGA, which is a population in its own right, forms a domain of study for planning and developmental purposes. In other words, it is advisable to treat certain parts of the population as strata if estimates are wanted separately for them.

Strata may also be formed solely for administrative convenience. The National Bureau of Statistics (NBS) has offices in all states of Nigeria, which collect data from each state. Here the states are treated as strata for administrative convenience.

In-Text Question

Strata may be form solely for administrative convenience. True or False

In-Text Answer

True.

5.3 Estimation in Stratified Sampling

The procedure of stratified sampling for the estimation of the population parameter is as follows:

(i) Divide the population of N units into L homogeneous non-overlapping strata of $N_1, \mathrm{N}_2, ..., N_L$ units respectively such that $N_1 + N_2 + ... + N_L = N.$

(ii) Select a sample of size $n_h, h = 1, 2, \cdots, L$ independently in the h^{th} stratum such that $n_1 + n_2 + \cdots + n_L = n,$ where n is the total sample size.

(iii) Calculate the required statistic in each stratum; these separate stratum statistics are then weighted to form the estimate of the characteristic of interest for the whole population.

From the foregoing, it is seen that the use of stratified sampling involves the following main operations:

(i) Choice of a stratification variable

(ii) Choice of the number of strata

(iii) Determination of the way in which the population is to be stratified

(iv) Choice of the stratum sample size

(v) Choice of sampling procedure in each stratum

It must be noted that classes of stratified sampling derive their names from the sampling schemes employed in drawing samples. But for the purpose of this course we shall be dealing with a class of stratified sampling called stratified random sampling. It is so called because a simple random sample is drawn independently from each stratum.

In-Text Question

The choice of the number of strata is one of the main operations of use of stratified sampling. True or False.

In-Text Answer

True

5.4 Estimation of Population Mean

In order to estimate the population mean, $\bar{Y}$ a simple random sample of size n_h is drawn without replacement from the population of N_h stratum units. In each stratum, information on the main character y is obtained from each unit that appears in the sample.

Let y_{hi} be the value obtained from the i^{th} sample unit in h^{th} stratum. The estimate of the population mean is given by

$$\bar{Y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \quad (st \text{, stands for stratified})$$

where

$$W_h = \frac{N_h}{N} = h^{th} \text{ Stratum weight}$$

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} = h^{th} \text{ stratum sample mean}$$

The population variance of $\bar{Y}_{st}$, for SRSWor is

$$V\left(\bar{Y}_{st}\right) = \sum_{h=1}^{L} W_h^2 \frac{1 - f_h}{n_h} S_h^2$$

where

$f_h = \frac{n_h}{N_h} = the \text{ h}^{\text{th}}$ stratum sampling fraction

S_h^2 = the h^{th} stratum population variance.

Note: The precision of $\bar{Y}_{st}$ depends on how far the stratum variability can be reduced.

The sample estimator of the population mean variance $V(\bar{Y}_{st})$ is

$$\hat{V}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n_h}$$

where

$$S_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{(n_h - 1)}$$

If interest is on the estimation of the population total, the estimator is

$$\hat{Y} = N\bar{y}_{st} = \sum_{h=1}^{L} N_h \, \bar{y}_h = \sum_{h=1}^{L} \hat{Y}_h$$

Its variance is

$$V(\hat{Y}_{st}) = \sum_{h=1}^{L} V(\hat{Y}_h) = \sum_{h=1}^{L} \frac{N_h (N_h - n_h)}{n_h} S_h^2$$

While the sample estimator of $V(\hat{Y}_{st})$ is

$$\hat{V}(\hat{y}_{st}) = \sum_{h=1}^{L} \frac{(N_h - n_n) S_h^2}{n_h} = N^2 \hat{V}(\bar{y}_{st})$$

5.5 Estimation of Population Proportion

For without replacement simple random sample of size n_h in h^{th} stratum, an unbiased estimator of the population proportion is $P_{st} = \sum_{h=1}^{L} W_h \, P_h$

P_h, is the sample proportion in h^{th} stratum.

The unbiased sample estimator of $V(P_{st})$ is

$$\hat{V}(P_{st}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{(N_h - 1)(n_h - 1)} P_h q_h = \sum_{h=1}^{L} W_h^2 \frac{1 - f_h}{n_h} \frac{N_h P_h q_h}{(N_h - 1)(n_h - 1)}$$

Example

1. All villages in a given region of a country were divided into 3 strata according to their land areas. A random sample of 26 villages was selected from within the strata and the cultivated area (in hectares) ascertained in each selected village. The results obtained are given below.

Stratum	N_h	**Cultivated area** y_{hi}
< 931	143	173, 476, 242, 281, 246, 195, 310, 531
931-1700	95	756, 634, 617, 553, 1022, 1187
>1700	40	839, 1091, 1158, 905, 788, 898, 980, 1119, 1823, 1238, 1068, 1331

Estimate the mean cultivated area per village and its standard error.

Estimate the total cultivated area in the region, and obtain the standard error of your estimate.

Solution

	Stratum			
	<931	931-1700	>1700	Total
N_h	143	95	40	**278**
n_h	8	6	12	
W_h	0.5144	0.3147	0.1439	**1.00**
$\overline{y}_h$	306.750	794.833	1103.167	
$W_h\overline{y}_h$	157.792	250.134	158.746	**566.672**
$N_h\overline{y}_h$	43,865.25	75,509.135	44,126.68	**163,501.065**
$1-f_h$	0.9441	0.9368	0.7	
S_h^2	16,838.214	64,576.567	78,048.879	
$W_h^2(1-f_h)\frac{S_h^2}{n_h}$	525.807	1177.228	94.277	**1797.312**
$N_h(N_h-n_h)\frac{S_h^2}{n_h}$	40632715.16	90999145.66	7284562.04	**138916422.9**

a) The mean cultivated area per village is given by the total in row seven of the table above. i.e.,

$$\overline{y}_{st} = 566.672$$

The estimate of the variance of $\overline{y}_{st}$ is given by the sum of the values in row eleven of the above table. i.e.

$\hat{V}(\overline{y}_{st}) = 1{,}797.312$ with standard error $Se = \sqrt{\hat{V}(\overline{y}_{st})} = 42.395$.

b) The total cultivated area in the region is

$\hat{Y}_{st} = N\bar{y}_{st} = 278 \times 566.672 = 157{,}534.816$ hectares. The estimate of the total cultivated area in the region is

$$\hat{V}(\hat{Y}_{st}) = N^2\hat{V}(\hat{Y}_{st}) = (278)^2(1{,}797.312) = 138{,}903{,}460.6$$
$$Se = \sqrt{N^2\hat{V}(\hat{Y}_{st})} = 11{,}785.731$$

5.6 Allocation of Sample Size to Strata

One of the problems in stratified sampling is the choice of the stratum sample size. In allocating sample sizes to strata the variability within stratum population size and the cost of obtaining information per element in each stratum must be taken into account.

However, if the variability and cost per element are not known in advance, a large sample may be selected from stratum that has a large stratum population size and a small sample from the stratum that has small stratum population size.

There are various methods of allocating sample size to strata. There are:

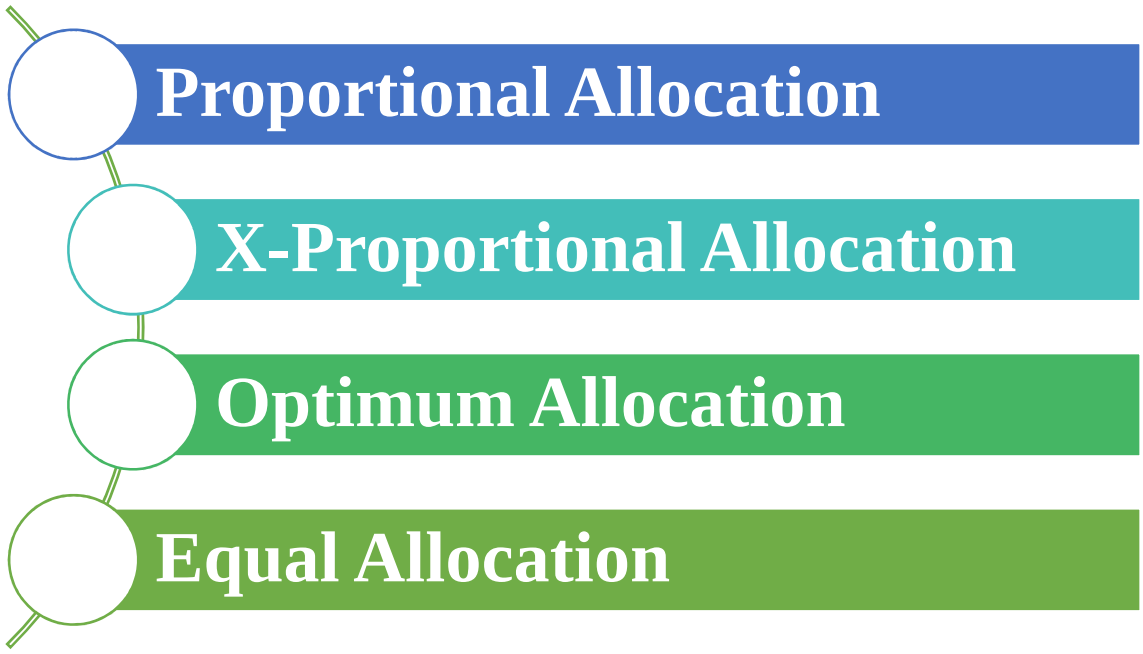


Figure5.2: Methods of allocating sample size to strata

5.5.1. Proportional Allocation:

In proportional allocation, also called N-proportional allocation because the sample allocation depends on the stratum size, the stratum sample is selected such that the size of the sample is proportional to the total number of units in each stratum. i.e.

$$n_h \propto \mathrm{N_h}$$
$$or$$
$$n_h \propto W_h$$

If the total sample size to be allocated is n, then stratum sample size is given as:

$$n_h = \frac{n}{N}\ \mathrm{N_h} = n\,\mathrm{W_h}$$

Thus, in proportional allocation

$$\frac{n_h}{n} = W_h$$
$$\Rightarrow \frac{n_h}{N_h} = \frac{n}{N} = f$$

in each stratum. This results in self-weighting sample.

If we substitute $\frac{n_h}{n}$ for W_h in

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h\ \bar{y}_h,$$

The result becomes

$$\bar{y}_{st} = \frac{1}{n}\sum_{h=1}^{L} n_h \bar{y}_h = \frac{1}{n}\sum_{h=1}^{L}\sum_{i=1}^{n_h} \bar{y}_{hi} = \bar{y}.$$

This shows that for proportional allocation, the sample mean $\bar{y}$ is the same as the stratified sample mean $\bar{y}_{st}$.

Variance in Proportional Allocation

Furthermore, if nW_h is substituted for n_h in

$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} S_h^2$,, the variance of $\bar{y}_{st}$ after simplification becomes

$$V_{pr}(\bar{y}_{st}) = \frac{1-f}{n} \sum_{h=1}^{L} W_h \, S_h^2$$

If the same substitution is made in the variance of the sample proportion given as

$$V(P_{st}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_n - 1} \frac{P_h Q_h}{n_h} = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} \quad \frac{N_h P_h Q_h}{(N_h - 1)}; \quad \textit{then}$$

variance reduces to

$$\mathrm{V}_{pr}(p_{st}) = \frac{1-f}{n} \sum_{h=1}^{L} \frac{W_h N_h P_h Q_h}{(N_h - 1)}$$

The gain made with proportional allocation depends on whether the variability within strata is less in stratum with smaller stratum size than in the larger stratum assuming that the cost of obtaining information from each unit is the same in all strata.

For practical purposes, proportional allocation is easy and simple to apply. It also yields modest gain in precision.

5.5.2 X-Proportional Allocation

In X-proportional allocation, the stratum sample size is made proportional to the total measure of size X_h in stratum h. This assumes that the measure of size x is known for each unit in the population. In this case the stratum sample size is

$$n_h = \frac{nX_h}{X}$$

Where X is the total measure of size for the whole population of N units.

The variance of $\bar{y}_{st}$ for X-proportional allocation becomes

$$V(\bar{y}_{st}) = \frac{1}{nN} \sum_{h=1}^{L} W_h \frac{\left(X - n\bar{X}_h\right)}{\bar{X}_h} S_h^2$$

The X-proportional allocation is preferable when sampling a skew population, for instance, Personal Incomes.

5.5.3 Optimum Allocation

The basic principle in optimum allocation is to allocate sample sizes to strata in such a way that large sample is taken from the stratum with high variability among its units or stratum with large stratum size; and small sample from the stratum with low variability among its units or stratum with small stratum size.

Or to increase the sample size in the stratum with low cost of obtaining information per unit and decrease the sample size in the stratum with high cost per unit. Optimum allocation is achieved when the stratum sample size is made proportional to the stratum standard deviation and inversely proportional to the square root of the cost per unit in the stratum; i.e.

$$n_h \propto \frac{S_h}{\sqrt{C_h}}$$

Hence, the problem of optimum allocation consists in minimizing the sampling variance for a given overall cost of the survey or minimizing the overall cost for specified sampling variance.

The optimum sample size is given as

$$n_h = \frac{\frac{n\,W_h S_h}{\sqrt{C_h}}}{\frac{\sum_{h=1}^{L} W_h S_h}{\sqrt{C_h}}} = \frac{\frac{n\,N_h S_h}{\sqrt{C_h}}}{\frac{\sum_{h=1}^{L} N_h S_h}{\sqrt{C_h}}}$$

Where C_h = cost per unit of obtaining the necessary information in h^{th} stratum.

If the cost per unit is the same in all strata $(C_h = C, h = 1, 2, ..., L)$,

$$n_h = \frac{n\,W_h\,S_h}{\sum_{h=1}^{L} W_h\,S_h} = \frac{n\,N_h\,S_h}{\sum_{h=1}^{L} N_h\,S_h}$$

This type of allocation where $C_h = C,$ in all strata is called Neyman allocation. It is a special case of optimum allocation

Variance in Optimum Allocation

$$V_{op}\left(\bar{y}_{st}\right)=\frac{1}{n}\left(\sum_{h=1}^{L}W_h\ S_h\ \sqrt{C_h}\right)\frac{\sum_{h=1}^{L}W_hS_h}{\sqrt{C_h}}-\frac{1}{N}\sum_{h=1}^{L}W_h\ S_h^2$$

where $\frac{\sum_{h=1}^{L}W_h\ S_h^2}{N}$ is the finite population correction $\left(fpc\right)$.

And for the Neyman allocation,

$$V_{ne}\left(\bar{y}_{st}\right)=\frac{1}{n}\ \left(\sum_{\mathrm{h=1}}^{\mathrm{L}}\mathrm{W_hS_h}\right)^2-fpc.$$

In-Text Question

The problem of optimum allocation consists in maximizing the overall cost for specified sampling variance. True or False

In-Text Answer

False

5.5.4 Equal Allocation

Another method of allocating samples to strata is by assigning equal sample sizes to all strata irrespective of the stratum population size, the stratum variability or cost per unit.

If *n* is the given total sample size, the stratum sample size is given by

$$n_h=\frac{n}{L}.$$

The variance of the stratified mean in equal allocation obtained by substituting $\frac{n}{L}$ for n_h is

$$V_{eq}\left(\bar{y}_{st}\right)=\frac{L}{n}\ \sum_{h=1}^{\mathrm{L}}\mathrm{W_h^2\ S_h^2}-\frac{\sum_{h=1}^{\mathrm{L}}\mathrm{W_h\ S_h^2}}{N}$$

Summary for Study Session 5

In study session 5, you have learnt that:

1. Stratification is a method of using auxiliary information to increase the precision of the estimate of population characteristic.
2. Strata may be formed because the sub-populations in which the entire population is sub-divided could each be treated as a population in its own right and designated as a domain of study
3. Classes of stratified sampling derive their names from the sampling schemes employed in drawing samples.
4. One of the problems in stratified sampling is the choice of the stratum sample size

Self-Assessment Questions (SAQs) for Study Session 5

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 5.1

Define stratification

SAQ 5.2

Outline the various methods of allocating sample size to strata

Notes on SAQ

SAQ 5.1

Stratification is a method of using auxiliary information to increase the precision of the estimate of population characteristic.

SAQ 5.2

- **Proportional Allocation**
- **X-Proportional Allocation**
- **Optimum Allocation**
- **Equal Allocation**

References

Cochran, W.G, (1977); *Sampling Techniques* third edition, New York: John Wiley & Sons

Daroga Singh and Chaudhary F.S, (1986); *Theory and Analysis of Sample Survey Design*, New Delhi: Wiley Eastern Limited

Des Raj and Promod Chandhok (1998); *Sample Survey Theory*, New Delhi: Narosa Publishing House

Kish L. (1965); *Survey Sampling*, New York: John Wiley & Sons

Okafor F.C (2002); *Sampling Survey Theory with Applications*, Nsukka: Afro-Orbis Publishers

Mukhopadhyay P. (2005): *Theory and Methods of Survey Sampling*, New Delhi: Prentice-Hall of India Private Limited

Study Session 6: Ratio and Regression Estimation

Introduction

The Survey sampler has been interested in methods of improving the precisions of estimates of population parameters both at the selection and estimation stages. If there is one thing that distinguishes sampling theory from general statistical theory, it is the degree of emphasis laid on the use of auxiliary information for improving the precision of estimates.

Auxiliary information was used in study session five for purposes of stratification. In this study, you will be introduced to some other methods of making use of auxiliary information to achieve higher precision.

Learning Outcomes for Study Session 6

At the end of this study, you should be able to:

6.1 Explain Ratio Estimation
6.2 Explain Regression Estimation

6.1 Ratio Estimation

Ration estimation is a technique that uses available auxiliary information which is connected with the variable of interest. In a sample survey, in addition to estimating the means, totals and proportions, you may wish to estimate the ratio of two characters.

Frequently, the quantity that is to be estimated from a simple random sample is the ratio of two variables both of which vary from unit to unit. Ratio estimates are prejudiced and modifications must be made when they are used in experimental or survey

For example, in a household survey the average number of suits of clothes per adult male, the average expenditure on cosmetics per adult female, and the average number of hours per week spent watching TV per child aged 10-15 years may be of interest. Or, the interest may

be in the estimation of income to expenditure, ratio of farmers' income to non-farmers' income, ratio of male to female enrolment in schools etc.

In order to estimate the first of these items, we would record for the ith household $(i = 1, 2, ...n)$ the number of adult males x_i who live in these households and the total number of suits y_i that they have. The population parameter to be estimated is the ratio R.

In-Text Question

Ration estimation is a technique that uses available auxiliary information. True or False

In-Text Answer

True

6.1.1 Definitions and Notations

Let $y_i =$ the value of the characteristic under study for the i^{th} unit of the population;

$x_i =$ The value of the auxiliary characteristic on the same unit;

$Y =$ The total of y characteristic of the population;

$X =$ The total of x characteristic of the population;

$R = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}} =$ The ratio of the population totals or means of character y or x

$\rho =$ The correlation coefficient between x and y in the population.

Suppose it is desired to estimate Y, or $\bar{Y}$ or R by drawing a simple random sample of n units from the population. Assume that based on n pairs of observations, $\bar{y}$ and $\bar{x}$ are the sample means of the characteristics y and x, respectively, and the population total X or mean $\bar{X}$ is known. The ratio estimators of the population ratio $R = \frac{Y}{X}$, the total Y, and the mean $\bar{Y}$, may be defined by:

$$\left.\begin{aligned} \hat{R} &= \frac{y}{x} = \frac{\overline{y}}{\overline{x}} \text{.....(6.2}i\text{)} \\ \hat{Y}_R &= \frac{\overline{y}}{\overline{x}} X = \hat{R}X \text{.....(6.2}ii\text{)} \\ \hat{\overline{Y}}_R &= \frac{\overline{y}}{\overline{x}} \overline{X} = \hat{R}\overline{X} \text{.....(6.2}iii\text{)} \end{aligned}\right\} \textit{respectively.}$$

Ratio estimator $\hat{R}$ unlike its components $\overline{y}$ and $\overline{x}$ is generally biased.

6.1.2 Approximate Variance of Ratio Estimator

Since ratio estimators are generally biased their mean square errors are considered for the purpose of comparing their efficiency with that of any other estimator. Ratio estimators, although biased, are consistent, and with simple random sampling for moderately large samples, the bias is negligible.

Consequently, for most practical purposes, approximate variance results are equally valid for comparison of its precision.

In simple random sampling, without replacement, if variates y_i and x_i are measured on each unit of a simple random sample of size n, assumed large, the Mean square error, (MSE) and variance of $\hat{R} = \frac{\overline{y}}{\overline{x}}$ are each approximately given by

$$MSE\ \left(\hat{\mathrm{R}}\right) = V\left(\hat{R}\right) = \frac{1-f}{n\overline{X}^2} \quad \frac{\sum_{i=1}^{N}\left(y_i - Rx_i\right)^2}{(N-1)}$$

Where

$R = \frac{\overline{Y}}{\overline{X}}$ is the ratio of the population means and

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

Proof:

$$\hat{R} - R = \frac{\overline{y}}{\overline{x}} - R = \frac{\overline{y} - R\overline{x}}{\overline{x}} \text{.....(6.2.2.0)}$$

If n is large, $\bar{x}$ should not differ greatly from $\bar{X}$. In order to avoid having to work out the distribution of the ratio of two random variables $(\bar{y} - R\bar{x})$ and $\bar{x}$, we replace $\bar{x}$ by $\bar{X}$ in the denominator of (6.2.2.0) as an approximation. This gives

$$\hat{R} - R = \frac{\bar{y} - R\bar{x}}{\bar{X}} \text{.....(6.2.2.1)}$$

Now, when we average over all simple random samples of size n,

$$\left.\begin{aligned} E\left(\hat{R} - R\right) &= \frac{E\left(\bar{y} - R\bar{x}\right)}{\bar{X}} \\ &= \frac{\bar{Y} - R\bar{X}}{\bar{X}} = 0 \end{aligned}\right\} \text{.....(6.2.2.2)}$$

$$R = \frac{\bar{Y}}{\bar{X}}.$$

From (6.2.2.1) we also obtain the result

$$MSE\left(\hat{R}\right) = E\left(\hat{R} - R\right)^2 = \frac{1}{\bar{X}^2} E\left(\bar{y} - R\bar{x}\right)^2 \text{.....(6.2.2.3)}$$

The quantity $(\bar{y} - R\bar{x})$ is the sample mean of the variate $d_i = y_i - Rx_i$, whose population mean $\bar{D} = \bar{Y} - R\bar{X} = 0.$

Note: The variance of the sample mean $\bar{y}$ is equal to the approximate variance of the ratio $\frac{\bar{y}}{\bar{x}}$ if the variate y_i is replaced by the variate $\frac{(y_i - Rx_i)}{\bar{X}}$.

A sample estimate of variance of the ratio estimator is given by

$$
\left.
\begin{aligned}
v(\hat{R}) &= \frac{(1-f)}{n\bar{X}^2}\frac{\sum_{i=1}^{n}\left(y_i-\hat{R}x_i\right)^2}{n-1} \\
&= \frac{(1-f)}{n\bar{X}^2(n-1)}\left[\sum_{i=1}^{n}y_i^2+\hat{R}^2\sum_{i=1}^{n}x_i^2-2\hat{R}\sum_{i=1}^{n}x_iy_i\right] \\
&= \frac{1-f}{n\bar{X}^2}\left(s_y^2+\hat{R}^2s_x^2-2\hat{R}s_{xy}\right) \\
&\textit{where} \\
&s_y^2 \ \&\ s_x^2 \text{ are the variances of X and Y respectively.} \\
&\mathrm{S_{xy}} = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{n-1}
\end{aligned}
\right\} \text{.....(6.2.2.5)}
$$

If the population mean of x is not known, its sample estimate may be used provided that the sample size is large $(n \geq 30)$.

For the estimated standard error of $\hat{R}$, this gives

$$
S\left(\hat{R}\right) = \frac{\sqrt{1-f}}{\sqrt{\mathrm{n}}\,\bar{X}}\sqrt{\frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}\left(\mathrm{y_i}-\hat{R}x_i\right)^2}{n-1}} \text{.....(6.2.2.6)}
$$

If $\bar{X}$ is not known, the sample estimate $\bar{x}$ is substituted in the denominator of (6.2.2.6).

One way to compute $S\left(\hat{R}\right)$ is to express it as

$$
S\left(\hat{R}\right) = \frac{\sqrt{1-f}}{\sqrt{n}\,\bar{x}}\sqrt{\frac{\sum y_i^2-2\hat{R}\sum y_ix_i+\hat{R}^2\sum x_i^2}{n-1}} \text{.....(6.2.2.7)}
$$

The sample estimator of the bias is given by

$$\left.\begin{aligned}\hat{B}(\hat{R}) &= \frac{1-f}{n\bar{X}^2}(\hat{R}s_x^2 - s_{xy}) = \frac{1-f}{n(n-1)\bar{X}^2}\left(\hat{R}\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}x_i y_i\right)\\ &\textit{Where}\\ s_{xy} &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}\end{aligned}\right\} \quad \text{.....(6.2.2.8)}$$

Example 6.1

A simple random sample (without replacement) of 20 villages was selected from the 88 villages in a given state. Using the sample observations on the area under maize cultivation y and land area of the village x collected from each of the sample villages in the table below:

y_i	242.82	245.65	352.90	755.57	247.68
x_i	313.39	764.05	655.27	797.72	310.80
y_i	553.22	792.81	609.48	1055.46	781.07
x_i	968.66	1134.42	1688.68	1623.93	1186.22
y_i	246.06	448.81	489.69	674.63	434.65
x_i	422.17	701.89	600.88	789.95	486.92
y_i	244.03	668.97	929.19	533.39	295.43
x_i	248.64	841.25	1440.04	1235.43	372.96

The total land area of the 88 villages is 83,819.96 hectares.

(i) Estimate the ratio of total area under maize cultivation to total land area in the state.

(ii) Estimate the total area under maize by method of ratio estimation.

(iii) Calculate the bias and variance of your estimates in (i) and (ii).

Solution

$N = 88$, $n = 20$, $\sum_{i=1}^{20} x_i = 16583.77$, $\bar{x} = 829.1885$, $\sum_{i=1}^{20} x_i^2 = 17,367,948$

$\sum_{i=1}^{20} y_i = 10601.51$, $\bar{y} = 530.0755$, $\sum_{i=1}^{20} y_i^2 = 6,794,275.214$;

$X = 83819.96; \bar{X} = 95.4995; f = \frac{20}{88} = 0.2273$

(i) Using (6.2i), the estimate of the ratio is

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{530.0755}{829.1885} = 0.639270202$$

(ii) The estimate of total area under maize cultivation is

$$\hat{Y}_R = \hat{R}X = 0.639270202 \times 83819.96 = 53,583.60 hectares$$

(iii) For the estimation of the variance and the bias, we shall use these sample results

$$s_y^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1} = \frac{6794275.214 - 20(530.0775)^2}{19} = 61,822.74178$$

$$s_x^2 = \frac{17367948 - 20(829.1885)^2}{19} = 190361.928$$

$$s_{xy} = \frac{10468440.11 - 20(530.0775)(829.1885)}{19} = 88302.988$$

The variance of $\hat{R}$ is calculated using

$$v(\hat{R}) = \frac{1-f}{n\bar{X}^2}\left(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}\right)$$

$$= \frac{1-0.2273}{20(952.4996)^2}\left[61822.742 + (0.6393)^2(190361.928) - 2(0.6393)(88302.988)\right]$$

$$= 0.00114$$

The estimate of the bias of R is

$$\hat{B}(\hat{R}) = \frac{1-0.2273}{20(19)(952.4995)^2}\left(0.6393\times 190361.928 - 88302.988\right)$$
$$= 0.00142$$

The ratio of the bias of $\hat{R}$ to its standard error is

$$\frac{\hat{B}(\hat{R})}{\sqrt{v(\hat{R})}} = \frac{0.00142}{0.03376} = 0.042$$

or

4.2%

For the variance of $\hat{Y}_R$, we use

$$\hat{V}(\hat{Y}_R) = X^2\hat{V}(\hat{R}) = (83189.96)^2(0.00114) = 8,009,212.228$$

In-Text Question

Ratio estimators are generally biased. True or False

In-Text Answer

True

6.1.3 Comparison with the simple average

The circumstances under which the **ratio estimate** will be better than the **simple average** (sample mean) will now be pointed out. The variance of $\hat{Y} = N\overline{y}$ in simple random sampling (without replacement) is

$$V(\hat{Y}) = N^2(1-f)\frac{S_y^2}{n} \text{.....(6.2.2.9)}$$

In this case no use is made of the auxiliary information provided by x. If this information is used to form the ratio estimate $\hat{Y} = X\hat{R}$, a first approximation to the mean square error around Y has been found to be

$$V_1(\hat{Y}) = N^2(1-f)\frac{(S_y^2 + R^2S_x^2 - 2R\rho S_x S_y)}{n} \quad \text{.....(6.2.2.10)}$$

Judging by this approximation, the ratio method will give a more precise result whenever

$$\left.\begin{array}{l} 2\rho > \dfrac{RS_x}{S_y} \\ or \\ \rho > \dfrac{C.V(x)}{2C.V(y)} \end{array}\right\} \quad \text{.....(6.2.2.11)}$$

Thus, the issue depends by and large on the strength of correlation between y and x. If x is the same character as y but has been measured on a previous occasion, the coefficient of variation may be taken to be equal. In that case it pays to use the ratio method of estimation if ρ exceeds 0.5. But one should not be dogmatic about the inequality (6.2.2.11) since it is based on an approximation. Ratio estimator is at its best when the relation between y and x is a straight line through the origin i.e. $y = kx$.

6.2 Regression Estimation

This is another method of using auxiliary information to improve the estimate of the population mean or total of the character under study. As has been noted, the ratio method is at its best when $y = kx$.

Conversely, the regression estimation is used to estimate the population mean when the regression line of y on x does not pass through the origin but makes an intercept along the $y-axis(y \neq kx)$.

6.2.1 Regression Estimation of Population Mean

Let $y_i, x_i (i = 1, 2, \cdots, n)$ be the sample values of the main character y and auxiliary character x respectively obtained on a simple random sample of size n selected without replacement from a population of size N. An estimator of the population mean of y is

$$\bar{y}_d = \bar{y} - k(\bar{x} - \bar{X}) \text{.....(6.2.2.12)}$$

k is a suitably chosen constant.

The estimator $\bar{y}_d$ is called difference estimator, which is unbiased for $\bar{Y}$.

Its variance is

$$V(\bar{y}_d) = V(\bar{y}) + k^2 V(\bar{x}) - 2kCov(\bar{x}, \bar{y})$$
$$= \frac{1-f}{n}(S_y^2 + k^2 S_x^2 - 2kS_{xy}) \quad \text{.....(6.2.2.13)}$$

To find the best value of k to use, we differentiate $V(\bar{y}_d)$ with respect to k and equate it to zero. This gives k optimum.

$k_{op} = \frac{S_{xy}}{S_x^2} = \beta$, the population regression coefficient.

The sample estimator of the $V(\bar{y}_d)$ is

$$\hat{V}(\bar{y}_d) = \frac{1-f}{n(n-1)} \sum_{i=1}^{n} \left\{ (y_i - \bar{y}) - k(x_i - \bar{x}) \right\}^2$$
$$= \frac{1-f}{n}\left(s_y^2 + k^2 s_x^2 - 2ks_{xy}\right) \quad \text{.....(6.2.2.14)}$$

When we substitute β, the regression coefficient, in place of k in (6.2.2.12) we have the estimator

$\bar{y}_d = \bar{y} - \beta(\bar{x} - \bar{X})$.....(6.2.2.15)

In practice β is not known but can be estimated from the sample at hand. The sample estimator of β is $\hat{\beta} = \frac{s_{xy}}{s_x^2}$.

Using this estimated value of β we have the linear regression estimator of the mean

$\bar{y}_{lr} = \bar{y} - \hat{\beta}(\bar{x} - \bar{X})$.....(6.2.2.16)

In large sample, the sample estimator of the variance of $\bar{y}_{lr}$ is

$\hat{V}(\bar{y}_{lr}) = \frac{1-f}{n}\left(s_y^2 + \hat{\beta}^2 s_x^2 - 2\hat{\beta}s_{xy}\right)$.....(6.2.2.17)

Or

$$\left.\begin{array}{l}\hat{V}(\bar{y}_{lr}) = \dfrac{1-f}{n} s_y^2 \left(1-\hat{\rho}^2\right) \\ \textit{where} \\ \hat{\rho}^2 = \dfrac{s_{xy}^2}{s_y^2 s_x^2}\end{array}\right\} \text{.....(6.2.2.18)}$$

Example 6.2

For the data in example 6.1, estimate the total area under maize cultivation using the method of regression estimation.

Solution

Since the population regression coefficient is not known, it is estimated from the sample data. From example 6.1

$$\hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{88302.982}{190361.93} = 0.46387$$

Using (6.2.2.16) the regression mean is

$$\bar{y}_{lr} = \bar{y} - \hat{\beta}(\bar{x} - \bar{X})$$
$$= 530.0755 - 0.46387(829.1885 - 952.4995) = 587.2758$$

The estimate of the total area under maize cultivation is

$$\hat{Y}_{lr} = N\bar{y}_{lr} = 88 \times 587.2758 = 51,680.268ha$$

The variance of this total is given by

$$\hat{V}(\hat{Y}_{lr}) = N^2 \hat{V}(\bar{y}_{lr})$$

Using (6.2.2.18) we need to obtain $\hat{\rho}^2$

$$\hat{\rho}^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{(88302.982)^2}{(190361.93)(61822.741)} = 0.66256$$

$$\hat{V}(\hat{Y}_{lr}) = N^2 \frac{1-f}{n} s_y^2 (1-\hat{\rho}^2)$$
$$= (88)^2 (0.03864)(61822.742)(0.33744)$$
$$= 6,242,338.104$$

In-Text Question

The ratio method is at its best when $y = kx$. True or False

In-Text Answer

True

Summary for study session 6

In study session 6, you have learnt that:

1. Ration estimation is a technique that uses available auxiliary information which is connected with the variable of interest. In a sample survey, in addition to estimating the means, totals and proportions, you may wish to estimate the ratio of two characters.
2. Ratio estimates are prejudiced and modifications must be made when they are used in experimental or survey.
3. Ratio estimators, although biased, are consistent, and with simple random sampling for moderately large samples, the bias is negligible.
4. Regression estimation is another method of using auxiliary information to improve the estimate of the population mean or total of the character under study

Self-Assessment Questions (SAQs) for Study Session 6

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 6.1

Define ratio estimation

SAQ 6.2

Explain the use of regression estimation

Notes on SAQ

SAQ 6.1

Ration estimation is a technique that uses available auxiliary information which is connected with the variable of interest

SAQ 6.2

This is another method of using auxiliary information to improve the estimate of the population mean or total of the character under study

References

Cochran, W.G, (1977); *Sampling Techniques* third edition, New York: John Wiley & Sons

Daroga Singh and Chaudhary F.S, (1986); *Theory and Analysis of Sample Survey Design*, New Delhi: Wiley Eastern Limited

Des Raj and Promod Chandhok (1998); *Sample Survey Theory*, New Delhi: Narosa Publishing House

Kish L. (1965); *Survey Sampling*, New York: John Wiley & Sons

Okafor F.C (2002); *Sampling Survey Theory with Applications*, Nsukka: Afro-Orbis Publishers

Mukhopadhyay P. (2005): *Theory and Methods of Survey Sampling*, New Delhi: Prentice-Hall of India Private Limited

Study Session 7:Non-Sampling Errors

Introduction

Apart from sampling error which you have learnt, there are other errors in a sample survey, which are not due to random sampling. These errors are called non-sampling errors. Non-sampling errors can occur even when information is canvassed from the whole population. They may occur anywhere from the planning stage to analysis stage of a survey.

In this study session, you will learn about non-sampling errors, the different types of non-sampling error, the sources of response error and the meaning of non-response error.

Learning outcomes for study Session 7

At the end of this study, you should be able to:

7.1 Define Non-sampling errors

7.2 Identify the different types of non-sampling error

7.3 Discuss the sources of response error

7.4 Explain non-response error.

7.1. Non-Sampling errors

A non-sampling error is a statistical error caused by human error to which an exact statistical analysis is revealed. Non-sampling errors are part of the total error that can result from doing a statistical analysis.

Non-sampling error is regarded as an important measure of the quality of data in addition to the sampling error, because it may distort seriously the result of the survey. Therefore, while minimizing the sampling error within the limits of available resources attempt should also be made to minimize the non-sampling error inherent in any survey.

Non-sampling error can occur either at the planning stage, execution stage or analysis stage, or at any combination of the stages.

The following are the possible causes of non- sampling errors at various stages of a survey.

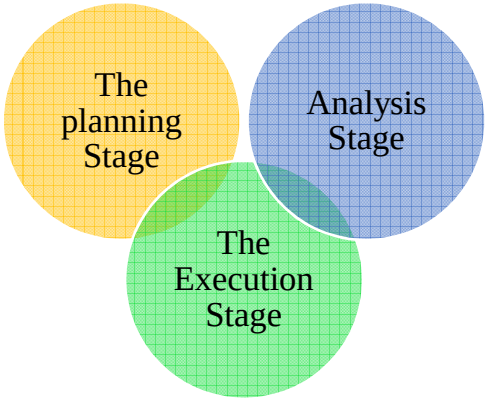


Figure 7.1: Non-Sampling errors

7.1.1 The Planning Stage:

Some factors that can cause non-sampling errors at the planning stage are:

1. Improper specification of the target population which may lead to over -coverage or under -coverage.
2. Wrong decisions on and/or ambiguous definitions of concepts to be used in preparing a frame and in designing a questionnaire may result in differing interpretations.
3. Omission or duplication of some units due to the use of obsolete frame.
4. Provision of inaccurate or faulty measuring instrument. Length of a questionnaire and inclusion of sensitive questions in a questionnaire may cause errors in a survey.

In-Text Question

Non-sampling error can occur on the following stages except

A. The planning stage
B. The execution stage
C. The analysis stage
D. The Descriptive stage

In-Text Answer

E. The Descriptive stage

7.1.2 The Execution Stage:

The Causes of errors at this stage include:

a. Inability of the enumerator to locate some of the units of enquiry, or inability to collect the required information from the units even when located.
b. Non-availability of experienced field investigators or inadequate training of field investigators.
c. Deliberate falsification of data by the enumerators or respondents.
d. Inability of the respondent to recall the required past information due to memory lapse (recall error) or his/her outright refusal to respond.

7.1.3 Analysis Stage:

Non-sampling error can occur at the analysis stage as a result of carelessness in editing the completed questionnaire. It can also be present during coding, data entry into the computer because of tiredness or carelessness. Also round-off error causes non-sampling error at this stage. Finally, errors may occur during tabulation and printing of final results due to tiredness or carelessness.

In-Text Question

Deliberate falsification of data by the enumerators or respondents can cause error at the execution stage. True or False

In-Text Answer

True

7.2 Types of Non-Sampling Error

Non-sampling error can be classified into two broad types namely:

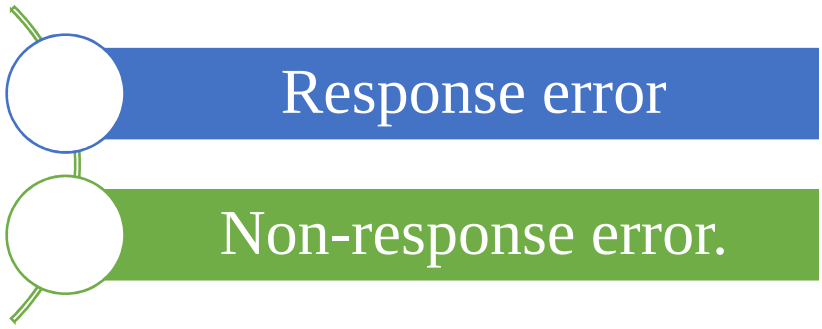


Figure 7.2: Classifications of non-sampling error

7.2.1 Response Error

Response error occurs when the respondent gives an incorrect value of desired characteristic. In practice, the value of the characteristic given by the i^{th} unit at interview, y_i may be different from the true value x_i. The difference between x_i and y_i is called the response error or observational error.

Consider the simple model

$$y_i = x_i + e_i$$
$$i = 1, 2, \cdots, N$$

If the response error, e_i is random then

$$\bar{Y} = \bar{X}$$

Since

$$\bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i = 0$$

In this case the true population mean value will be the same as the observed mean value. If on the other hand the error is systematic, i.e. tilted to one direction, then

$$\bar{Y} - \bar{X} = \bar{e}$$

Where $\bar{e}$ is the response bias which does not cancel out. And the true population mean will differ from the observed mean value.

7.3 Sources of Response Error

There are several sources of response error. They are:

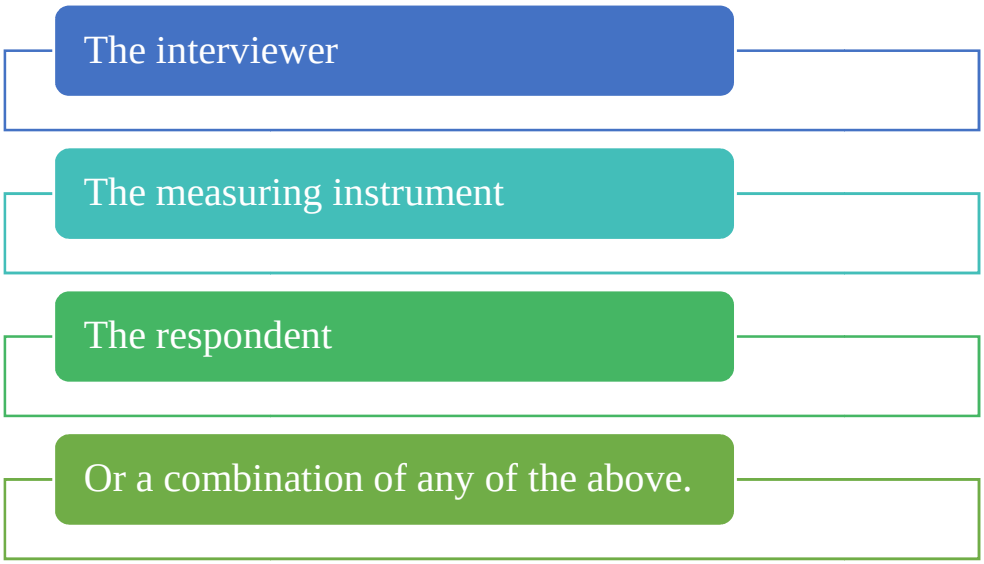


Figure 7.3: Sources of response error

7.3.1 The interviewer

The interviewer may fail to record or code correctly the answers supplied by the respondent due to carelessness or fatigue. He may ask questions in a different way from the original wordings of the questions.

This may mislead the respondent and thus lead him/her into giving inaccurate answer. The interviewer due to lack of interest in the subject matter of the survey or in the interviewing process itself may cook up data without actually interviewing any respondent.

Inadequate training of the interviewers and lack of proper supervision of the fieldwork are also factors that can cause response error. The manner and extent of probing by the interviewer in order to elicit more information from the respondents may give room for response error. Finally, interviewer's personality and his/her manner of interaction with the respondent is another source of response error.

7.3.2 Instrument

Faulty measuring instruments like weighing balance, questionnaire or schedule may give rise to false responses. Length of the questionnaire may cause fatigue thus reducing concentration and this may lead to inaccurate responses.

Figure 7.4: Instrument-weighing balance

Questions dealing on the prestige of the respondent, or on socially unacceptable behaviour or questions that centre on the intimacy of the respondent may result in false answers. When answers to questions that are too technical in nature are sought from a novice, you may not be surprised to have many wrong answers due to lack of respondent's understanding of the terms used. Also ambiguous and sensitive questions or leading words in a questionnaire may produce false answers.

7.3.3 Respondent

Answering questions without the respondent quite clearly understanding the questions asked may result in the respondent giving the wrong answers. The respondent may, on certain occasion, deliberately give ego-boosting answers even when he is fully aware that the answer is incorrect.

At other time, the respondent, out of ignorance of the nature or use of the survey, may deliberately give wrong answer.E.g. In a rural survey of household income and expenditure, the respondent for fear of taxation may fail to disclose all his farms or give a lower income. Thus, the attitude of the respondents towards the survey usually affects the survey results.

Another case of response error on the part of the respondent is memory lapse or recall error which depends on the length of the reporting period and the date of the survey. Usually, question that are too sensitive, such as sexual behaviour, when asked directly are full of response errors, especially when asked in the presence of a third party.

Information collected from any available adult in a household survey is more likely to cause response error, since only the head of the household can supply some household information accurately.

In-Text Question

Memory lapse can be viewed as response error. True or False

In-Text Answer

True

7.4 Non-Response Error

Non-response error results from failure to obtain responses to questions from some of the units in a census or survey. Non-response lately is been used to measure the quality of data since it introduce bias in the estimates. In addition, the size of the non-response gives indication of the reliability of survey data.

Non-response can be classified into three categories namely:

Non-coverage, item non- response and total (unit) non- response.

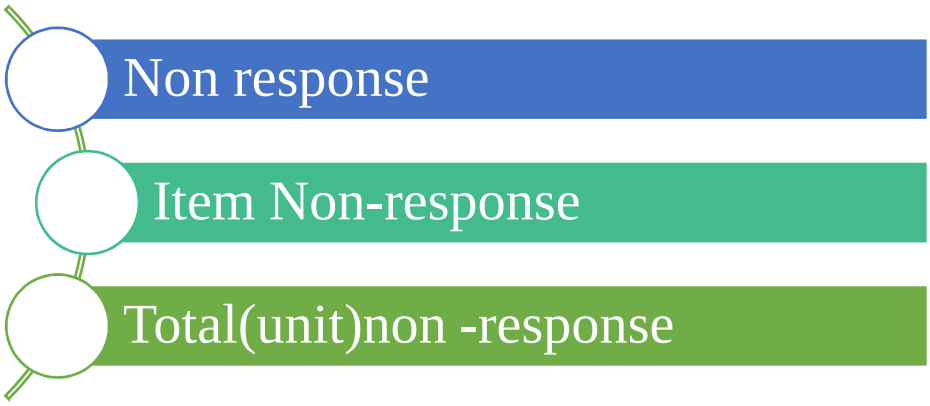


Figure 7.5: Classifications of non-response error

I. Non- response

Non- response refers to failure to include some units, which properly belong to the target survey population in the sampling frame. This is not, properly speaking, a non-response problem.

Ii. Item Non-response

Item non-response occurs as a result of inability to obtain response to one or more but not all questions in the questionnaire from a maple reporting unit.

Causes

The causes include the following:

1. It is caused by the respondent's inability or refusal to answer a particular question in the questionnaire.

2. It may reoccur during editing of returned completed questionnaires when inconsistent answers are deleted.

3. Item non-response is also caused by the failure of the interviewer to ask a particular question or when he or she fails to record an answer supplied by the respondent.

Iii Total (Unit) non-response

Total or unit non-response occurs when none of the responses to questions in the questionnaire are obtained from a reporting unit in the sample.

Causes

Total non- response arises because of the following:

1. **Not-at-home:** The respondents may not be at home at the time the interviewer calls. Another category are those who may not be chanced to grant interview at the time the interviewer calls but are willing to cooperate at another convenient time. The two categories require call back.

2. **Movers/Not Found: These** are those who are in the sample but who changed residence at the time of the survey. It includes those who cannot be reached because of hostile environment or inaccessibility as a result of floods or bad terrain, etc.

3. **Refusals**: These are the sample units who are unwilling to cooperate. These people refuse to grant interview even after several appeals. They are the hard-core.

4. Away from home: There are those sample persons or households who are away from home for one reason or the other throughout the duration of the survey.

5. **Lost Questionnaire:** This occurs when the collected information is misplaced. The completed questionnaires may be lost while in transit to the office or lost in the office itself through carelessness. Lost questionnaire includes those rejected as a result of poor quality or because the information was collected from the wrong person.

6. **Unsuitable for Interview:** Some sample units may be unable to participate in the survey because of serious illness and others because of language barrier.

In-Text Question

_____occurs when none of the responses to questions in the questionnaire are obtained from a reporting unit in the sample.

A. Total or unit non-response
B. Item response
C. Non response
D. Item –non response

In-Text Answer

A. Total or unit non-response

Summary for study session 7

In study session 7, you have learnt that:

1. A non-sampling error is a statistical error caused by human error to which an exact statistical analysis is revealed. Non-sampling errors are part of the total error that can result from doing a statistical analysis.
2. Non-sampling error is regarded as an important measure of the quality of data in addition to the sampling error, because it may distort seriously the result of the survey.

3. Non-sampling error can occur at the analysis stage as a result of carelessness in editing the completed questionnaire. It can also be present during coding, data entry into the computer because of tiredness or carelessness
4. Response error occurs when the respondent gives an incorrect value of desired characteristic. In practice, the value of the characteristic given by the i^{th} unit at interview, y_i may be different from the true value x_i .
5. Non-response error results from failure to obtain responses to questions from some of the units in a census or survey. Non-response has of late been used to measure the quality of data since it introduce bias in the estimates.

Self-Assessment Questions (SAQs) for Study Session 7

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 7.1 (Testing Learning outcomes 7.1)

Define non sampling error

SAQ 7.2 (Testing Learning outcomes 7.2)

Identify the different types on non-sampling errors

SAQ 7.3 (Testing Learning outcomes 7.3)

Outline the sources of response error.

SAQ 7.4 (Testing Learning outcomes 7.4)

Define non response error.

Notes on SAQ

SAQ 7.1

A non-sampling error is a statistical error caused by human error to which an exact statistical analysis is revealed. Non-sampling errors are part of the total error that can result from doing a statistical analysis.

SAQ 7.2

- Response error
- Non-response error.

SAQ 7.3

- The interviewer
- The measuring instrument
- The respondent
- Or a combination of any of the above.

SAQ 7.4

Non-response error results from failure to obtain responses to questions from some of the units in a census or survey. Non-response lately is been used to measure the quality of data since it introduce bias in the estimates.

References

Cochran, W.G, (1977); *Sampling Techniques* third edition, New York: John Wiley & Sons

Daroga Singh and Chaudhary F.S, (1986); *Theory and Analysis of Sample Survey Design*, New Delhi: Wiley Eastern Limited

Des Raj and Promod Chandhok (1998); *Sample Survey Theory*, New Delhi: Narosa Publishing House

Kish L. (1965); *Survey Sampling*, New York: John Wiley & Sons

Okafor F.C (2002); *Sampling Survey Theory with Applications*, Nsukka: Afro-Orbis Publishers

Mukhopadhyay P. (2005): *Theory and Methods of Survey Sampling*, New Delhi: Prentice-Hall of India Private Limited