# MODERN SURVEY SAMPLING

ARIJIT CHAUDHURI

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

# MODERN SURVEY SAMPLING

# MODERN SURVEY SAMPLING

ARIJIT CHAUDHURI

INDIAN STATISTICAL INSTITUTE,
KOLKATA, INDIA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742


Version Date: 20140520

International Standard Book Number-13: 978-1-4665-7261-4 (eBook - PDF)

# *Dedication*

*To*
*Bulu*

# Contents

# List of Tables

# Preface

At last this is my venture to address a textbook on sample surveys to an international readership.

This text is meant for students taking undergraduate-level university courses in statistics and also for those taking courses at the graduate and master levels in statistics.

The present author also wrote an elementary-level textbook called *Essentials of Survey Sampling* published by Prentice Hall of India in New Delhi in January 2010, which is supposed to be printed as a revised second edition in early 2014.

*Survey Sampling: Theory & Methods* is by the present author in collaboration with Professor H. Stenger of Mannheim University, Germany, with the first edition published in 1992 by Marcel Dekker, New York and a thoroughly revised and enhanced second edition published by Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, FL in 2005.

All three texts present a sizable amount of material on various aspects of survey sampling. Yet, there is a need for what may be considered an international-level textbook on survey sampling despite there being quite a few publications available in the international market on the present subject.

As I have been teaching this subject for numerous years at the Indian Statistical Institute and have considerable experience with this topic, having published profusely in numerous journals, I wish to share with those around the world my work and knowledge gained as a veteran teacher.

Applied Statistics Unit
Indian Statistical Institute,
203, B.T. Road,
Kolkata - 700 108, India

Arijit Chaudhuri
email: arijitchaudhuri1@rediffmail.com
November 2013

# Acknowledgments

I am grateful to the Director of the Indian Statistical Institute for giving me an opportunity to continue to work so many years since retirement to teach and publish. Also, I appreciate the congenial atmosphere created by my colleagues in the Applied Statistics Division of the Institute for productive pursuits.

# About the Author

**Arijit Chaudhuri** is Honorary Visiting Professor at the Indian Statistical Institute (ISI) after serving as a CSIR Emeritus Scientist and earlier as a Professor in ISI. He has a Ph.D. in Statistics from Calcutta University. He is an author of *Randomized Response and Indirect Questioning Techniques in Surveys* (Chapman & Hall, CRC Press, 2011), *Essentials of Survey Sampling* (Prentice Hall of India, 2010), and co-author of *Indirect Questioning in Sample Surveys* jointly with TC Christofides (Springer-Verlag, 2013), *Survey Sampling Theory and Methods* (1992 1st. edition, Marcel Dekker and 2nd. edition jointly with H. Stenger, Chapman & Hall, CRC Press, 2005), *Randomized Response: Theory & Techniques* (Marcel Dekker, 1988) jointly with Rahul Mukherjee, and *Unified Theory and Strategies of Survey Sampling* (North-Holland, 1988) jointly with late JWE Vos. The author has widely travelled with academic assignments in universities and statistical offices in the United States, Canada, England, Australia, Sweden, Germany, the Netherlands, South Africa, Japan, Turkey, Cuba, Cyprus, and Israel. He is the founder-chairman of a registered "Advanced Survey Research Centre" (website: www.asrc.net.in).

Arijit Chaudhuri
email: arijitchaudhuri1@rediffmail.com

# 1 Exposure to Sampling

Abstract. Introduction. Concepts of population, sample, and sampling.

## 1.0 ABSTRACT

A survey means organized observation with the purpose of reaching conclusions in a scientific manner. Observation relates to a totality, called a Population or a Universe, and a part thereof is a Sample. It is possible and useful to survey a sample to make inferences concerning parameters that mean characteristics of a population. In survey sampling a parameter typically is an unknown real number. Using observed real values for a selected sample, a derived real number is proposed as a possible value of the unknown parametric value or as a point estimator. As an alternative, an interval with this point estimator within it along with two numbers on either side of it is claimed to contain within itself the unknowable parametric value with a reasonable claim for such an assertion to hold true. This is Interval Estimation. An organizer of a Sample Survey is assigned a task to explain how to choose a sample appropriately to justify the specification from the observed sample values a point estimator and an interval estimator in a way acceptable to a scientific community.

## 1.1 INTRODUCTION

The dominating topic in Survey Sampling is estimation. A real-life problem demanding practical application is here the motivating factor. But to formulate, develop, and study a related theory, certain abstractions are naturally needed. We stumble first upon the conceptualization of what is called a Population or a Universe. This denotes the totality of all objects of interest in a given context. For example, all the building structures on a street in Kolkata constitute a Population. So, all the 50 states plus the National Capital Territory, Washington, D.C., form a universe in the context of all the components of the United States. A serious study of certain features of the constituent components of such a Population would be a tremendous task. So, handling only a few of these parts of a Population should appear reasonable while being cognizant of the general characteristics of all the elements of such Populations. To grasp an essential idea in such a situation, it is judged imperative to hit upon the concept of a Sample which, nontechnically speaking, is but a Part of a Population. For the individuals in the sample, respective values of one or more variables of interest are ascertained to the extent possible. Some suitable functions of the values on the sampled individuals are taken as suitable Statistics,

combined judiciously, if necessary, with other available variate-values for the sampled units, or all the population members are employed to estimate the values of parameters representing characteristics of interest defined for the population of all the individuals concerned. If a statistic is worked out as a real number to be the value of a parameter of interest defined as a real number, then the statistic is treated as a point estimator, for the real-valued parameter. Sometimes an interval around the value of this point estimator taking some sampled variate-values suitably combined with certain other suitable constants, is constructed, claiming to contain within itself the parameter value with a high probability. This interval is called a Confidence Interval. The probability associated as above with such an interval is called the Confidence Coefficient.

The problem to be addressed by the Investigator is how to choose an appropriate sample and procedures for Point and Interval estimation for real-valued parameters defined on Populations of interest.

## 1.2 CONCEPTS OF POPULATION, SAMPLE, AND SAMPLING

By $U = (1, \ldots, i, \ldots, N)$ we denote a Population of a known number $N$ of individuals also called members, units, or elements, in the context of Sample Surveys or Survey Sampling. Each element of $U$, say, $i$, is supposed to be identifiable and assigned labels as $i$ for identification and referencing. For a population of all the villages in a given district in a province in a country, Names will be the identifier but will be substituted as label $i$, which bears the values $1, 2, 3, \ldots$ etc.

The crop-fields in a given village in their turn must also be tagged with such labels of positive integers for identification and referrals, though thus tagged every element is a tangible, concrete object in a Survey Population which is finite with a known number $N$ of objects. If we consider a pond in a city or town or a village the number of fishes in it is of course finite at every instant of time, but this number for all practical purposes must be considered as unknown. So, the theory we are going to establish cannot cover such a concept as a Population of fishes in a given pond. This is because no individual fish in a pond can be identified and tagged without causing damage to its life and liberty. However, taking special care for such contingencies, a modified theory will be described in brief to cover such populations which are finite but composed of unknown and unidentifiable entities.

Treating $i_1, i_2, \ldots, i_n$ each as one of the labels $i$ in $U$, we shall denote by the sequence $(i_1, \ldots, i_j, \ldots, i_n) = s$, a Sample from $U$. Here, the order in which the labels occur in $s$ is important and such an $s$ is called an "ordered" sample from $U$. These labels in $s$ need not all be "distinct." Yet number of labels in $s$ is recognized as $n$, and this $n$ is called the "size" of the Sample $s$. But by $v(s)$ we denote the number of Distinct units in $s$, and this $v(s)$ is called the "Effective Size" of the sample $s$. Of course, $1 <= v(s) <= n$. By Sampling we mean the act of selecting a sample from the population. In order that a scientific theory may be developed for sample selection and estimation of

Population Parameters using the samples drawn and surveyed, it is useful if one chooses a sample $s$ with a pre-assigned probability, denoted, say, as $p(s)$. Since it is a probability, of necessity, we must have

$$\text{(i) } 0 <= p(s) <= 1$$
$$\text{(ii) } \sum_{s} p(s) = 1.$$

Here $\sum_{s}$ denotes summing over all possible choices of a sample $s$ from $U$.

A sample is more usefully defined as a set $s^* = \{i_1, \ldots, i_n\}$ of distinct labels $i_1, i_2, \ldots, i_n$ which are the $n$ distinct entities of $U$. Here it is immaterial in which order the respective labels of $U$ are written in $s^*$. Here $n$ is taken as the Size of the sample $s^*$ and it is obviously its effective size as well. Here $\binom{N}{n}$ is the total number of possible such unordered samples of $n$ distinct units drawable from the Population $U$ of $N$ units. By $p(s^*)$ we mean the selection-probability of $s^*$ from $U$, and we need

$$\text{(i) } 0 <= p(s^*) <= 1$$
$$\text{(ii) } \sum_{s^*} p(s^*) = 1.$$

Here $\sum_{s^*}$ denotes sum over all the $\binom{N}{n}$ possible number of ways defining the samples like $s^*$ that may be chosen from $U$.

Any such function $p$ defined on the totality of all possible samples like $s$ or $s^*$ from $U$ described above with the two specified properties is called a Sampling Design. In practice, for simplicity, we shall write $s$ to denote either a Sequence-type or a Set-type sample avoiding the cumbrous symbol $s^*$ unless it is crucial to stress that we mean to imply an "unordered" sample of only "distinct units."

We shall throughout, unless emphasized otherwise, mean to use a sample $s$ to be chosen with a certain probability $p(s)$ employing a design (more elaborately a sampling design) $p$. Thus, $s$ is a random variable.

In surveying a sample our concern will be to observe the values $y_i$ for a main real variable $y$ of interest for the respective units $i$ in the sample $s$ actually chosen to be surveyed. Of course the real values $y_i$ are defined for every $i(= 1, 2, \ldots, N)$ in the population $U$ defining the vector $\underline{Y} = (y_1, \ldots, y_i, \ldots, y_N)$. Likewise, other real variables $x, z, w$, etc., are with respective values $x_i, z_i, w_i$, etc., for $i$ in $U$ and the vectors

$$\underline{X} = (x_1, \ldots, x_i, \ldots, x_N),$$
$$\underline{Z} = (z_1, \ldots, z_i, \ldots, z_N),$$
$$\underline{W} = (w_1, \ldots, w_i, \ldots, w_N).$$

The prime interest in Sample Surveys is to suitably estimate the population total $Y = \sum_1^N y_i$ of $y$ using the survey data denoted $d = (s, y_i | i \in s)$ along with knowledge of $\underline{Z}$, $\underline{W}$, etc., plus partial knowledge at least of $\underline{X}$ as, for example, the value of $X = \sum_1^N x_i$ being prima facie known. Let $t' = t(d) = t(s, \underline{Y})$ with the restriction that $t(s, \underline{Y})$ does not involve any $y_i$ in $\underline{Y}$ unless $i$ is in $s$ (i.e., it is free of $y_j$ for $j \notin s$).

Such a function $t$ of $d$ is called a statistic. Appreciating that $\underline{Y}$ is a vector of fixed but unknown values $y_i$ for $i \in s$, this $t(s, \underline{Y})$ is a random variable because it is a function of the random variable $s$ even though the other component in $t(s, \underline{Y})$ is a constant. So, we may define

$$E_p(t) = \sum_s p(s) t(s, \underline{Y})$$

as the expectation of $t$ with respect to the design $p$ which provides $t(s, \underline{Y})$ its multiplier in $E_p(t)$.

Such a function $t$ of $t(s, \underline{Y})$ may be employed to estimate $Y$. The unknowable value of $t - Y = t(s, \underline{Y}) - Y$ is called the error in estimating $Y$ by the value of $t(s, \underline{Y})$ for a sample $s$ at hand. The expected value of this error viz.

$$E_p(t - Y) = B_p(t) = \sum_s p(s)(t(s, \underline{Y}) - Y)$$

is defined and called the Bias of $t$ in estimating $Y$ using the data

$$d = (s, y_i | i \in s)$$

on choosing the sample $s$ on implementing the design $p$. The square error $(t - Y)^2$ has the expectation

$$E_p(t - Y)^2 = \sum_s p(s)(t(s, \underline{Y}) - Y)^2$$

called the Mean Square Error (MSE) of $t$ in estimating $Y$.

Again $$V_p(t) = \sigma^2 = \sigma_p^2(t) = E_p(t - E_p(t))^2$$

is defined. This is called the variance of $t$ in respect of the design $p$ which has given the probability $p(s)$ to the sample $s$ to be selected for being surveyed, yielding the data $d$ and the estimator $t$ for $Y$.

Clearly, $$\text{MSE} = V_p(t) + B_p^2(t)$$

i.e., $$\text{MSE} = \text{Variance} + \text{Squared Bias}.$$

The estimator $t$ for $Y$ is called a Point Estimator for $Y$, as it is just a value that is a real number proposed to represent the value $Y$ which is just an unknown real number. The performance characteristics of $t$ as a Point Estimator for

$Y$ are the quantities $E_p(t)$, $B_p(t)$, $M_p(t)$, and $V_p(t) = \sigma_p^2(t)$. The quantity $\sigma_p(t) = +\sqrt{V_p(t)}$ is called the Standard Error of $t$.

Let us consider the expanded quantity

$$M_p(t) = \sum_s p(s)(t-Y)^2 = \sum_1 p(t)(t(s,\underline{Y})-Y)^2 + \sum_2 p(s)(t(s,\underline{Y})-Y)^2,$$

denoting by $\sum_1$ the sum over the samples $s$ for which $|(t(s,\underline{Y})-Y)|$ exceeds a positive number $K$, briefly

$$\sum_1 = \sum_{s:|t(s,\underline{Y})-Y|\geq K>0} \text{ and } \sum_2$$

denoting the sum over the samples $s$ in the complementary set so that

$$\sum_2 = \sum_{s:|t(s,\underline{Y})-Y|<K} .$$

Then it follows that

$$MSE \geq K^2 \sum_{s:|t(s,\underline{Y})-Y|\geq K} = K^2 \text{Prob}[|t(s,\underline{Y})-Y| \geq K]$$

writing Prob[.] for the probability of the event denoted by the symbol[.].

Hence, it follows that

$$\text{Prob}[|t(s,\underline{Y})-Y| \geq K] \leq \frac{E_p(t-Y)^2}{K^2}$$

or

$$\text{Prob}[|t(s,\underline{Y})-Y| \geq K] \leq \frac{{\sigma_p}^2(t)+{B_p}^2(t)}{K^2}$$

Choosing a positive number $\lambda$ such that $K = \lambda\sigma_p(t)$ it follows that

$$\text{Prob}[|t-Y| \geq \lambda\sigma_p(t)] \leq \frac{1}{\lambda^2} + \frac{1}{\lambda^2}(\frac{(B_p(t))}{\sigma_p(t)})^2$$

or

$$\text{Prob}[t-\lambda\sigma_p(t) \leq Y \leq t+\lambda\sigma_p(t)] \geq (1-\frac{1}{\lambda^2}) - \frac{1}{\lambda^2}(\frac{(B_p(t))}{\sigma_p(t)})^2$$

So, whatever may be the vector $\underline{Y}$ of real numbers $y_i, i \in U$, it follows that the random interval

$$CI = (t-\lambda\sigma_p(t), t+\lambda\sigma_p(t))$$

contains the unknown number $Y = \sum_1^N y_i$ within itself with a probability at least as high as

$$(1-\frac{1}{\lambda^2}) - \frac{1}{\lambda^2}(\frac{(B_p(t))}{\sigma_p(t)})^2 = CC$$

This interval CI is called a Confidence Interval covering the parameter $Y$ within itself with a Confidence Coefficient at least as high as this number CC. Reporting this about this CI in terms of CC is called Interval Estimation of $Y$. The situation simplifies greatly for an estimator $t$ for $Y$ ensuring $B_p(t) = 0$ for every $\underline{Y}$. Such a $t$ is called an unbiased estimator for $Y$. In such a case $(t - \lambda\sigma_p(t), t + \lambda\sigma_p(t))$ provides a Confidence Interval for $Y$ with a Confidence Coefficient at least as high as $(1 - \frac{1}{\lambda^2})$.

The quantity $2\lambda\sigma_p$ gives the width of the Confidence Interval. It is desirable to have a Confidence Interval with a small width. Thus it is desirable to employ for $Y$ an unbiased point estimator $t$ with $\sigma_p(t)$ small in magnitude for any prescribed choice of $\lambda$ for which $(1 - \frac{1}{\lambda^2})$ should be as large as, say, equal to 0.95 or 0.99, giving us a CI with a CC at least as high as 95 or 99. The problem of sampling then for an investigator to solve is stipulating a sampling design $p$, throwing up the survey data $d$ and a point estimator which is unbiased for $Y$, admitting a small $\sigma_p(t)$ so as to provide an accurate estimation rule, and as a by product yielding a CI with a desirably small width and a high Confidence Coefficient.

# 2 Initial Ramifications

Abstract. Introduction. Sampling design, sampling scheme. Random numbers and their uses in random sampling (SRS). Drawing simple random samples with and without replacement. Estimation of mean, total, ratio of totals/means: variance and variance estimation. Determination of sample sizes. Appendix to Chapter 2.

## 2.0 ABSTRACT

Only a few rudimentary concepts related to survey sampling are briefly set forth in this chapter. How to select a sample for which a selection-probability is specified to prescribe its performance characteristics is to be clearly narrated. Concepts of random samples and simple random samples are to be laid bare. How many samples are to be chosen and with what requirements also need to be clarified. How to measure accuracy in terms of unknowable features and how to assess its realization through measurements are briefly discussed.

## 2.1 INTRODUCTION

In Chapter 1 we noted that a finite Survey Population $U = (1, 2, \ldots, i, \ldots, N)$ containing $N$ distinctly identifiable units labeled, respectively, as $1, 2, \ldots, i, \ldots, N$ is said to have $N$ as its size which is a finite positive integer known to the investigator. A sample $s$ from $U$ may be either a sequence of a finite number of labels ordered successively as the first, second, etc., to the $n$th element $i_1, i_2, \ldots, i_n$ each of which is one of the labels of the units of $U$. Thus, $s = (i_1, \ldots, i_j, \ldots, i_n)$, and $n$ is the size of this ordered sample and the elements in this sequence $s$ need not all be distinct. Alternatively a sample $s$ from $U$ may denote a set of $n$ distinct units of $U$ with no regard for the order of succession in which these labels are inserted in the sample $s$ which is a set of $n$ distinct units of $U$. This number $n$ is the size of the sample $s$. Here $1 \leq n \leq N$. In case a sample $s$ is a sequence of ordered units $n$ in number, not all of which need to be distinct from one another, this sample $s$, though it has the size as $n$, also has an Effective Size which is the number of distinct labels out of these total of $n$ labels that are contained in $s$. How to draw samples with what rationales and how to use them with what purpose are issues to be settled.

## 2.2 SAMPLING DESIGN, SAMPLING SCHEME

By Sampling Design we mean a probability measure $p$ that assigns to a sample $s$ a selection-probability $p(s)$ with two properties as discussed in Chapter 1. Its principal role is to guide us in developing a unified theory of estimating from survey data Population parameters of theoretical and practical interest.

In contrast a Sampling Scheme specifies methods of actual selection of samples. When so laid down, a sampling scheme prescribes probabilities of selection of samples. Thus, a sampling scheme develops a sampling design. Correspondingly, given a sampling design, it is possible to work out a sampling scheme. Let us see some details.

If for every possible sample $s$ from $U$ there is given a procedure for its selection with a given probability $p(s)$, say, then a sampling design $p$ is already formed. To grasp the idea of its converse also being true—that is, given a sampling design $p$ assigning to a sample $s$ its selection-probability $p(s)$—let us refer to Hanurav's (1962) classical device which gives us a corresponding procedure of choosing such a sample according to an actual scheme of selection by a one-by-one draw of units from the population.

Let $s = (i_1, \ldots, i_j, \ldots, i_n)$ denote a sample of which $i_j$ is the unit chosen from $U$ assigned the $j$th label for $j = 1, \ldots, n$. Let $p(s)$ values for each such $s$ be given. Hanurav (1962) gives a draw-by-draw procedure for its selection.

Let $p(i_1)$ = probability of choosing the singleton sample $(i_1)$,

$p(i_1, i_2)$ = probability of choosing the sample $(i_1, i_2)$, etc., and finally,

$p(i_1, \ldots, i_j, \ldots, i_n)$ = given probability of choosing $s = (i_1, \ldots, i_j, \ldots, i_n)$

Let $\alpha_{i_1} = \sum_1 p(s)$, with $\sum_1$ as the sum over all samples with $i_1$ as its first element; $\alpha_{i_1 i_2} = \sum_2 p(s)$ with $\sum_2$ as the sum over all samples of which $i_1, i_2$ are the first two units and so on; and finally $\alpha_{i_1 i_2 \ldots i_n} = \sum_n p(s)$ with $\sum_n$ as the sum over all samples with $i_1, i_2, \ldots, i_n$ as the first $n$ units in the samples. Then Hanurav's (1962) sampling scheme specifies the following:

1. Make the first draw from $U$ with probability $\alpha_{i_1}$ to get $i_1$ as the first unit in the sample $s$.
2. Then, implement a Bernoullian trial with $\left(1 - \frac{\beta_{i_1}}{\alpha_{i_1}}\right)$ as the probability of "success" stipulating to make another draw only on realizing a "success," stopping further exercise in case of a "failure."
3. Then, if a 'success' results draw the unit $i_2$ with probability $\frac{\alpha_{i_1 i_2}}{\alpha_{i_1} - \beta_{i_1}}$.
4. Next perform another Bernoulli trial with probability of success $\left(1 - \frac{\beta_{i_1 i_2}}{\alpha_{i_1 i_2}}\right)$ with a similar stipulation as in step (2). If a failure results in step (4), one ends up getting the sample $(i_1, i_2)$ with probability

$$\alpha_{i_1} \left(\frac{\alpha_{i_1} - \beta_{i_1}}{\alpha_{i_1}}\right) \frac{\alpha_{i_1 i_2}}{\alpha_{i_1} - \beta_{i_1}} \frac{\beta_{i_1 i_2}}{\alpha_{i_1 i_2}} = \beta_{i_1 i_2}.$$

With similar additional steps one chooses following this procedure a sample $s = (i_1, \ldots, i_j \ldots, i_n)$ with probability

$$\alpha_{i_1} \left(\frac{\alpha_{i_1} - \beta_{i_1}}{\alpha i_1}\right) \times \frac{\alpha i_1 i_2}{\alpha_{i_1} - \beta_{i_1}} \times \cdots \times \frac{\beta_{i_1 i_2 \cdots i_n}}{\alpha_{i_1 \cdots i_n}} = \beta_{i_1 i_2 \cdots i_n}.$$