SURVEY SAMPLING

PARIMAL MUKHOPADHYAY

α

Alpha Science

SURVEY SAMPLING

Parimal Mukhopadhyay

Alpha Science International Ltd.
Oxford, U.K.

Parimal Mukhopadhyay
Professor Indian Statistical Institute (Retd.)
203 B.T. Road, Kolkata, India

Alpha Science International Ltd.
7200 The Quorum, Oxford Business Park North
Garsington Road, Oxford OX4 2JZ, U.K.

Printed from the camera-ready copy provided by the Author.

ISBN-13: 978-1-84265-374-6
ISBN-10: 1-84265-374-1

Printed in India

SURVEY SAMPLING

Dedicated to the memory of my uncle

Late Tripureshwar Mukhopadhyay

Preface

This monograph makes a comprehensive review of some of the developments in the theory of survey sampling in the superpopulation model-based approach, starting from an assessment of the situation in the classical fixed-population area.

Chapters 1 and 2 are based on the model of fixed population. Chapter 1 introduces the preliminary concepts, sampling designs, estimators, sampling strategies, various classical estimators, etc. Chapter 2 addresses some inferential problems, e.g., uniformly minimum variance unbiased estimation, admissibility, sufficiency, minimax estimation in survey sampling. The remaining chapters have developed from the assumption of some superpopulation models depicting the survey population.

Chapter 3 considers model-dependent optimal strategies in the prediction-theoretic approach. Chapter 4 deals with the robustness of these strategies under specific model-failures; Chapter 5, the class of strategies which combine randomization both due to sampling designs and the superpopulation models. Chapter 6 addresses the asymptotic properties of these strategies and their robustness.

The following chapter examines the robustness of model-dependent and modelbased strategies in the asymptotic sense. This chapter also identifies regression superpopulation models for which the prediction-estimators become robust. Biasrobust estimation including non-parametric calibration are also discussed. In the design-based conditional approach of Chapter 8, inference is restricted to a part of the sample space which satisfies certain properties. Conditionally optimum estimators are studied in this light. The next chapter addresses design-based calibration estimation of finite population parameters under different distance functions. The concepts of calibration with restricted weights, extended calibration, mitigated calibration are examined. Model-based calibration is discussed in the following chapter and its optimality investigated. The concluding chapter considers yet another approach to estimation, empirical maximum likelihood estimation in finite population sampling. The performance of this approach vis-a-vis calibration approach is examined.

As has been noted in Chapter 2, the arguments based on a fixed-population model do not lead to any optimality result in general. In a broader perspective, survey population can be looked upon as a realization of a superpopulation and many decisive results can be attained in this set-up. The thrust of the book is, therefore, on superpopulation model-based inference. One important finding is that the generalized regression estimator (*greg*) plays a prominent part, specially, in large scale sample surveys.

This book does not cover, among others, Bayesian approaches in survey sampling, model-based variance estimation and an important application, small area estimation. These have been covered in some details in Mukhopadhyay (2000a, 1996, 1998b) respectively.

The book, to some extent, may be considered as an up-to-date version of, but not restricted to, works in Mukhopadhyay (1996). However, many useful results of the earlier book have not been revisited here.

In writing this book I have attempted to arrange and reconcile the results systematically in a lucid manner and indicate new research areas. Various examples and supplementary exercises have been added to clarify the ideas. We have assumed that the reader has a basic degree in Statistics and is acquainted with the developments in survey sampling at the level of Cassel, *et al.* and Mukhopadhyay (1998a). The book can not be a stand-alone text book for the fixed-population part, but may serve as a self-contained study material for the model-based part, generally of use to the researchers.

The book was partially written during my assignment in the University of South Africa, Pretoria. My family helped me a lot by silent inspiration. An acknowledgement is due to my daughters-in-law Jayita and Shilpi who assisted me in arranging the manuscripts.

Kolkata, India

Parimal Mukhopadhyay

Contents

Chapter 1

The Preliminaries

1.1 Introduction

Sample survey, finite population sampling or survey sampling is a method of drawing inference about the characteristic of a finite population by observing only a part of the population. Different statistical techniques have been developed to cater to these needs during the last few decades.

In this chapter we examine a model for a fixed finite population which formed the basis of the earlier exposition of the theory. The set-up established in this chapter would be fundamental to our discussion mainly up to Chapter 2.

1.2 The Basic Model

We assume that we have a finite population of distinct units, with a known population size and a variable of interest taking real values on these units. In an enumerative survey the primary task of the survey statistician is to estimate some descriptive characteristics of the population, e.g., population total, mean, variance by suitably choosing a subset (sample) of the population and observing the values of this variable only on the units in the selected subset. (In analytic surveys we consider estimation of superpopulation parameters and these will be considered in Chapter 3 onwards. For the time being we consider the fixed population model and the associated enumerative surveys).

To formulate the basic fixed population model precisely let us consider a few definitions.

DEFINTION 1.2.1: A finite (survey) population $\mathcal{P}$ is a collection of a known number N

N of identifiable units labelled $1, \ldots, i, \ldots, N, \mathcal{P} = \{1, \ldots, i, \ldots, N\}$, where i stands for the physical unit labeled i.

The above definition excludes from its coverage the populations of the following types: batches of industrial products of the same specification coming out from a production process as the units are not distinguishable individually; population of fishes in a lake as the population size is unknown. Collection of households in an area, industrial units in an urban complex, agricultural fields in a village are examples of survey populations.

Let y be a study variable having value y_i on $i = 1, \ldots, N$. Associated with $\mathcal{P}$ we have a vector $\mathbf{y} = (y_1, \ldots, y_N)$ which constitutes the parameter for the model of a survey population, $\mathbf{y} \in \mathcal{R}^N$, the parameter space. One is often interested in estimating a parameter function $\theta(\mathbf{y})$, e.g., population total, $T = \sum_{i=1}^{N} y_i$, population mean, $\bar{y} = \sum_{i=1}^{N} y_i/N = T/N$, population variance $S^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_i - \bar{y})^2$ by choosing a sample (a part of the population, defined below) from $\mathcal{P}$ and observing the values of y only on the units in the sample.

DEFINITION 1.2.2: A sample is a part of the population.

A sample may be selected with replacement (*wr*) or without replacement (*wor*) of the units already selected to the original population.

A sample when selected by a *wr* -sampling procedure may be written as a sequence,

$$S = \{i_1, \ldots, i_n\},\ 1 \leq i_t \leq N,\ t = 1, \leq n, \tag{1.2.1}$$

where i_t denotes the label of the unit selected at the tth draw and is not necessarily equal to $i_{t'}$ for $t \neq t'(= 1, \ldots, N)$. For a without replacement sampling procedure, a sample when written as a sequence, is

$$S = \{i_1, \ldots, i_n\},\ 1 \leq i_t \leq N, i_t \neq i_{t'} \text{ for } t \neq t'(= 1, \ldots, N) \tag{1.2.2}$$

since repetition of units in S is not possible. Arranging the units in the sample S in an increasing (decreasing) order of magnitudes of labels and considering only the distinct units, a sample may also be written as a set s. For a *wr*-sampling of n draws, a sample written as a set is, therefore,

$$s = (j_1, \ldots, j_{\nu(S)}),\ 1 \leq j_1 < \ldots < j_{\nu(S)} \leq N \tag{1.2.3}$$

where $\nu(S)$ is the number of distinct units in S. In a *wor*-sampling procedure, a sample of n draws, written as a set is,

$$s = (j_1, \ldots, j_n),\ 1 \leq j_1 < \ldots < j_n \leq N. \tag{1.2.4}$$

Thus, if in a *wr*-sampling, $S = \{2, 5, 2, 1\}$, the corresponding s is $s = (1, 2, 5)$ with $\nu(S) = 3$. Similarly, if for a *wor*-sampling procedure, $S = \{3, 7, 1\}$, the corresponding s is $s = (1, 3, 7)$. Clearly, information on the order of selection and repetition of units in the sample S is not available in s.

DEFINITION 1.2.3: Number of distinct units in a sample is its effective sample size. In (1.2.3), $\nu(S)$ is the effective sample size, $1 \leq \nu(S) \leq n$. For a *wor*-sample of n draws, $\nu(S) = \nu(s) = n$.

Note that a sample is a sequence or set of some units from the population and does not include their y-values.

DEFINITION 1.2.4: The sample space is the collection of all possible samples and is often denoted as $\mathcal{S}$. Thus $\mathcal{S} = \{S\}$ or $\{s\}$ according as we are interested in S or s.

In a simple random sample with replacement (*srswr*) of n draws $\mathcal{S}$ contains N^n samples S. In a simple random sample without replacement (*srswor*) of n draws $\mathcal{S}$ contains $(N)_n$ samples S and $\binom{N}{n}$ samples s where $(a)_b = a(a-1)\dots(a-b+1), a > b$. If the samples s of all possible sizes are considered in a *wor*-sampling procedure, there are 2^N samples in $\mathcal{S}$.

DEFINITION 1.2.5: Let $\mathcal{A}$ be the minimal σ-field over $\mathcal{S}$ and p a probability measure defined over $\mathcal{A}$ such that $p(s)$ [or $p(S)$] denotes the probability of selecting s[or S], satisfying

$$\begin{aligned} p(s)\ [p(S)] &\geq 0 \\ \textstyle\sum_{s\in\mathcal{S}} p(s)\ [\sum_{S\in\mathcal{S}} p(S)] &= 1. \end{aligned} \tag{1.2.5}$$

One of the main tasks of the survey statistician is to find a suitable $p(s)$ or $p(S)$. The collection $(\mathcal{S}, p)$ is called a sampling design, often denoted as $D(\mathcal{S}, p)$ or simply p. The triplet $(\mathcal{S}, \mathcal{A}, p)$ is the probability space for the model of the finite population.

The expected effective sample size of a sampling design p is

$$E\{\nu(S)\} = \sum_{S\in\mathcal{S}} \nu(S)p(S) = \sum_{\mu=1}^{N} \mu P[\nu(S) = \mu] = \nu. \tag{1.2.6}$$

We shall denote by ρ_n the class of all fixed effective size $[FES(\nu)]$ designs, i.e.

$$\rho_n = \{p : p(s) > 0 \Rightarrow \nu(S) = \nu\}. \tag{1.2.7}$$

A sampling design p is said to be noninformative if $p(s)[p(S)]$ does not depend on the y-values. In this treatise, unless stated otherwise, we shall consider non-informative designs only. Informative designs have been considered by Basu (1969), Zacks (1969), Liao and Sedransk (1975), Stenger (1977), Bethlehem and Schuerhoff (1984), among others.

Basu (1958), Basu and Ghosh (1967) proved that all the information relevant to making inference about the population characteristic is contained in the set sample s and the corresponding y-values (Theorem 2.3.1). As such, unless otherwise stated, we shall consider samples as sets s only.

The quantities

$$\pi_i = \sum_{s \ni i} p(s), \quad \pi_{ij} = \sum_{s \ni (i,j)} p(s)$$

$$\pi_{i_1 \dots i_k} = \sum_{s \ni (i, \dots, i_k)} p(s) \tag{1.2.8}$$

are, respectively, the first order, second order, ..., kth order inclusion-probabilities of units in a sampling design p. The following lemma depicts some relations among inclusion-probabilities and expected effective sample size of a sampling design.

Lemma 1.2.1: For any sampling design p,
(i)

$$\pi_i + \pi_j - 1 \le \pi_{ij} \le \min(\pi_i, \pi_j)$$

(ii)

$$\sum_{i=1}^{N} \pi_i = \sum_{s \in S} \nu(s) p(s) = \nu$$

(iii)

$$\sum \sum_{i \neq j = 1}^{N} \pi_{ij} = \nu(\nu - 1) + V\{\nu(s)\}.$$

If $p \in \rho_n$,
(iv)

$$\sum_{j(\neq i)=1}^{N} \pi_{ij} = (\nu - 1)\pi_i$$

(v)

$$\sum \sum_{i \neq j = 1}^{N} \pi_{ij} = \nu(\nu - 1).$$

Result (i) is obvious. Results (ii), (iii) and (iv), (v) are, respectively, due to Godambe (1955), Hanurav (1962), Yates and Grundy (1953).

Further, for any sampling design p,

$$\theta(1 - \theta) \le V\{\nu(s)\} \le (N - \nu)(\nu - 1) \tag{1.2.9}$$

where $\nu = [\nu] + \theta, 0 \le \theta < 1, \theta$ being the fractional part of ν. The lower bound in (1.2.9) is attained by a sampling design for which

$$P[\nu(S) = [\nu]] = 1 - \theta \text{ and } P[\nu(S) = \nu + 1] = \theta.$$

Mukhopadhyay (1975) derived a sampling design with fixed nominal sample size $n(> \nu), [p(S) > 0 \Rightarrow n(S) = n \ \forall S]$ such that $V\{\nu(S)\} = \theta(1 - \theta/(n - [\nu])$, which is very close to the lower bound in (1.2.9). Here, $n(S)$ is the nominal sample size (number of draws in) S.

We shall denote by

$p_r(i_r)$ = probability of selecting i_r at the rth draw

$p_r(i_r|i_1,\ldots,i_{r-1})$ = conditional probability of selecting i_r at the rth draw given that $i_1,\ldots,i_{r-1}$ are selected at the first, ..., $(r-1)$th draws respectively.

Suppose the values $x_1,\ldots,x_N$ of a closely related (to y) auxiliary variable x on units $1,2,\ldots,N$ respectively, are available. As an example, in an agricultural survey, x may be the area of a plot under a specified crop and y the yield of crop on that plot. The quantities $p_i = x_i/X, X = \sum_{i=1}^{N} x_i$ is called the size-measure of unit $i(=1,\ldots,N)$ and is often used in selection of samples.

1.3 Different Types of Sampling Designs

Sampling designs proposed in the literature can be generally grouped in the following categories:

(a) Simple random sampling with replacement (*srswr*)

(b) Simple random sampling without replacement (*srswor*)

(c) Probability proportional to size with replacement (*ppswr*) sampling: a unit i is selected with probability p_i at the rth draw and a unit once selected is returned to the population before the next draw ($r = 1,2,\ldots$).

(d) Unequal probability without replacement (*upwor*) sampling: A unit i is selected at the rth draw with probability proportional to $p_i^{(r)}$ and a unit once selected is removed from the population. Here

$$p_1(i) = p_i^{(1)}$$

$$p_r^{(i_r)} = \frac{p_{i_r}^{(r)}}{1 - p_{i_1}^{(r)} - p_{i_2}^{(r)} - \ldots - p_{i_{r-1}}^{(r)}}, r = 1,2,\ldots,n. \tag{1.3.1}$$

The quantities $\{p_i^{(r)}\}$ are generally functions of p_i and the p_i-values of the units already selected. In particular, if $p_i^{(r)} = p_i \ \forall \ i = 1,\ldots,N$, the procedure may be called *probability proportional to size without replacement* (*ppswor*) sampling procedure. For $n = 2$, for this procedure

$$\pi_i = p_i[1 + A - \frac{p_i}{1-p_i}]$$

$$\pi_{ij} = p_i p_j [\frac{1}{1-p_i} + \frac{1}{1-p_j}], \quad \text{where } A = \sum_{k=1}^{N} \frac{p_k}{1-p_k}.$$

The sampling design may also be attained by an inverse sampling procedure where units are drawn wr, with probability $p_i^{(r)}$ at the rth draw, until for the first time n distinct units occur. The n distinct units each taken only once constitute the sample.

(e) Rejective sampling: Drwas are made wr and with probability $\{p_i^{(r)}\}$ at the rth draw. If all the units turn out distinct, the solution is taken as a sample; otherwise, the whole sample is rejected and fresh dras are made. In some situations $p_i^{(r)} = p_i \ \forall \ i$.

(f) Systematic sampling with varying probability (including equal probability).

(g) Sampling from groups: The population is divided into L groups either at random or following some suitable procedures and a sample of size n_h is drwan from the hth group by using any of the above-mentioned sampling dsigns such that the desired sample size $n = \sum_{h=1}^{L} n_h$ is attained. An example is the Rao-Hartley-Cochran (1962) sampling procedure.

Based on the above methods, there are many uni-stage or multi-stage stratified sampling procedures.

A $FES(n)$-sampling design with π_i proportional to p_i is often used for estimating a population total. This is, because an important estimator, the Horviyz-Thompson estimator (HTE) has zero variance if y_i is proportional to p_i. Such a design is called a πps design or *IPPS* (inclusion-probability proportional to size) design. Since $\pi_i \leq 1$, it is required that $x_i \leq X/n \ \forall \ i$ for such a design.

Many (exceeding sixty) unequal probability without replacement sampling designs have been suggested in the literature, mostly for use along with the HTE. For many of these designs sample size is a variable. Again, some of these designs are sequential in nature (e.g., Chao (1982), Sunter (1977)). Mukhopadhyay (1972), Sinha (1973), Herzel (1986) considered the problem of realizing a sampling design with pre-assigned sets of inclusion-probabilities of first two orders.

In a sample survey, all the possible samples are not generally equally preferable from the view-point of practical advantages. In agricultural surveys, for example, the investigators tend to avoid grids which are located further away from the cell camps, are located in marshy land, inaccessible places, etc. In such cases, the sampler would like to use only a fraction of totality of all possible samples, allotting only a small probability to the non-preferred units. Such designs are called *Controlled Sampling Designs* and have been considered by several authors (e.g., Chakravorty (1963), Srivastava and Saleh (1985), Rao and Nigam (1989, 1990), Mukhopadhyay and Vijayan (1996)).