Chapter 1

Introduction

1.1 Survey sampling: a historical perspective

Among various branches of statistics as it is known today, survey sampling is the one which has an origin in the prehistoric human societies. Probably its use had started by checking whether a cluster of apples or some other fruits ripened properly or not by plucking and tasting one or a few apples from the cluster. Later, when human beings started cultivation in one form or the other this tasting or checking habit became one of the routine practices. As Jeffreys (1961) says in another similar context,

The process is so habitual that we hardly notice it and we can hardly exist for a minute without carrying it out. On the rare occasions when anybody mentions it, it is called common sense and left at that (Jeffreys 1961, page 1).

This informal use of survey sampling continued for quite a long time. But unlike other branches, surprisingly, its major objective has been the same from prehistoric to today’s modern societies. The objective of any formal or informal use of survey sampling is to gain knowledge about a particular characteristic possessed by each individual unit in a population having finitely many units by checking only a small part of the population. The objective remains the same whether we are interested in knowing the quality of fruits or grains or today’s opinion polls and market surveys.
Perhaps the formal theory of survey sampling started its journey with Neyman (1934). This work was particularly stimulated by Bowley (1906). Smith (1976) gave an elegant account of Neyman's work and its influence on practitioners as well as theoreticians. The nonparametric approach adopted by Neyman attracted practising statisticians to a large extent. However, a large part of theoretical statisticians shied away from this theory till fifties.

The work on formalising the theory of survey sampling was taken up around the same time by Indian statisticians also. This was mainly in the field of estimating the agricultural yields. Prominent among these statisticians were Mahalanobis (1940, 1946), Panse (1944) and Sukhatme (1945).

Another fundamental contribution, which nevertheless was a logical follow-up of Neyman (1934), was Hurvitz and Thompson (1952). This paper is well known for its introduction of varying probability sampling. It also provided the impetus for new research at foundational level, such as the concept of linear estimators.

It may be said that the voyage of modern survey sampling began with Godambe (1955). Now this paper is well recognized for developing a unified theory of survey sampling and for proving the non-existence of a uniformly minimum variance unbiased estimator even among the class of linear estimators. As Smith (1976) puts it, this was just a beginning of doubts. Rather turbulent times were ahead. If the uniformly minimum variance approach had failed, did it imply that other approaches such as sufficiency, likelihood, admissibility and so on would also fail? Most of these questions were answered by Godambe and others.

For instance, it was proved by Basu (1958) that the set of all distinct labels in the sample is a sufficient statistic. As a consequence, any estimator can be \textit{Rao-Blackwellized} (for a precise meaning of this term we refer to Cassel, Särndal and Wretman 1977, page 39). It is a well known fact that with sufficiency alone we are unable to go beyond this.
A similar fate was faced by the concept of admissibility developed by Godambe and Joshi in a series of papers beginning in 1966. However, situation was slightly better for the decision theoretic analysis in 1980's. This was mainly due to a new technique developed by Hsuan (1979) and Meeden and Ghosh (1983) for proving admissibility, termed stepwise Bayes procedure. This will be discussed in detail in Chapter 2.

Among all these, it was the likelihood approach which put Neyman's theory in a defensive position. It was proved by Godambe (1966) that if we adopt a likelihood approach, then any estimation procedure based on the likelihood function has to be free of the the sampling design. This was rather embarrassing because sampling designs are the building blocks of Neyman's frequentist theory. Also, by the likelihood approach we can say virtually nothing about the unobserved part of the population and everything about the observed part of the population. This can be understood in a simpler way. Since there are no assumptions about the underlying population characteristic values, it is natural that if the population consists of non-similar objects, we can not make any conclusion about the unsampled part by looking at the sampled part.

It was at this juncture that Basu (1971) and Ericson (1969a) wholeheartedly recommended use of Bayesian analysis in survey sampling. Not only that it implements the likelihood principle in a rather mechanical way, Bayesian method has strong philosophical underpinning. Before developing the techniques of Bayesian analysis in the context of survey sampling, we give a brief discussion on Bayesian philosophy in the next section.

The above discussion is restricted to design based inference. This was in the sense that the probabilities refered only to those introduced by the design used for selecting the sample. Thus the interpretation to the optimal procedures were with respect to repeated sampling. The values of the characteristics for the population
units were assumed to be fixed though unknown. Later on the probability models for population characteristics have also been considered. This is what we now call as the superpopulation approach. Classical inference using such superpopulation models is considered by several researchers, prominent among them being Brewer (1963) and Royall (1970). Such an approach indeed is still design based. A good account of such work is found in three books: Bolfarine and Zacks (1992), Cassel, Särndal and Wretman (1977) and Särndal, Swensson and Wretman (1991).

1.2 Bayesian reasoning

Bayesian reasoning is based on the famous Bayes theorem, which we state as follows: For any propositions \( H \) and \( E \),

\[
\text{Prob}(H \mid E) = \frac{\text{Prob}(E \mid H)\text{Prob}(H)}{\text{Prob}(E)}.
\]

In the usual applications of the theorem, \( H \) is some hypothesis and \( E \) the evidence against which it is to be evaluated. \( \text{Prob}(H) \) is the amount of belief, or prior probability attached to the hypothesis \( H \); \( \text{Prob}(E \mid H) \) is the likelihood of \( H \) on \( E \). \( \text{Prob}(H \mid E) \) is the amount of belief in \( H \) after observing the evidence \( E \). \( \text{Prob}(H \mid E) \) is called the posterior probability of \( H \) after observing \( E \). The theorem can be rewritten as

\[
\text{Prob}(H \mid E) \propto \text{Prob}(E \mid H)\text{Prob}(H)
\]

i.e. posterior probability \( \propto \) prior probability \( \times \) likelihood.

\( \text{Prob}(E \mid H) \) can be interpreted as the degree with which \( E \) explains \( H \) and \( \text{Prob}(H) \) is the prior estimate of the weight of evidence in favour of \( H \) before \( E \) becomes
known. In practice it is by balancing empirical and prior factors in just this sort of way that hypotheses are evaluated. Neither factor by itself is decisive.

What are the consequences of the theorem in scientific reasoning in general and, statistical inference in particular? This is well debated. The theorem precisely tells us how a rational mind should behave. We refer to elegant discussions by Edwards, Lindman and Savage (1963), Howson and Urbach (1991) or Lindley (1990).

Another aspect well appreciated by the theorem is the predictive nature of statistical inference. To make this precise, let us reformulate the theorem in the following manner: Let \( y_n = (y_1, \ldots, y_n)^T \) be a vector of sample observations and \( \theta \) a vector of parameters, then,

\[
p(\theta \mid y_n) = \frac{p(\theta)p(y_n \mid \theta)}{p(y_n)}.
\]

Here \( p(y_n \mid \theta) \) can be interpreted as the distribution of \( y_n \) or the likelihood of \( \theta \), depending on whether we are looking at it as a function of \( y_n \) or \( \theta \) respectively. The probability distribution \( p(\theta) \) represents the accumulated prior knowledge before the observation on \( y_n \) is made. Now compute the posterior distribution \( p(\theta \mid y_n) \) and if a future observation \( y_{n+1} \) comes from \( p(y_{n+1} \mid \theta) \), we compute

\[
p(y_{n+1} \mid y_n) = \int_{\theta} p(\theta \mid y_n)p(y_{n+1} \mid \theta)d\theta.
\]

We can base our probabilistic prediction of a future observation \( y_{n+1} \) on \( y_n \) and prior \( p(\theta) \) through the predictive distribution \( p(y_{n+1} \mid y_n) \). Now suppose that the future is similar to the past in the sense that

for every \( n \), the distribution \( p(y_1, \ldots, y_n) \) is the same as \( p(y_{i(1)}, \ldots, y_{i(n)}) \)

for every permutation \( i(1), \ldots, i(n) \) of the first \( n \) positive integers

then de Finetti's theorem guarantees the existence of a parameter \( \theta \) with prior \( p(\theta) \) that characterizes this similarity. In other words, given \( \theta \), \( y_i \)'s are independent and
identical observations from $p(. | \theta)$ and marginally

$$p(y_n) = \int \prod_{i=1}^{n} p(y_i | \theta)p(\theta)d\theta$$

where $p(\theta)$ is a prior for $\theta$. This notion of similarity is known as exchangeability.

However, de Finetti's concept of exchangeability is not directly applicable in a finite population context, as it is for an infinite sequence $\{y_1, y_2, \ldots\}$. What we want is finite exchangeability.

We say that a finite sequence $y_1, \ldots, y_N$ is finite exchangeable if the distribution $p(y_1, \ldots, y_N)$ of $y = (y_1, \ldots, y_N)^T$ is the same as $p(y_{i(1)}, \ldots, y_{i(N)})$ for every permutation $i(1), \ldots, i(N)$ of the first $N$ positive integers, where $N$ is fixed.

We refer to Diaconis (1981) for a theoretical discussion on finite exchangeability and its relationship with the de Finetti's theorem. Throughout the thesis, by exchangeability we mean finite exchangeability.

The concept of exchangeability is rather fundamental in the context of survey sampling. However, it is popularly known as homogeneity within a population. The concept of a homogeneous population belongs to the folklore of survey sampling theory; it is coupled with the notion of a representative sample. Whenever a survey practitioner talks about a representative sample, which is the selection of part of an aggregate to represent the whole, the idea of a homogeneous population is implicit. Even though there are differences from a finer point of view (cf. Lindley and Novick 1981), we treat homogeneity and exchangeability as one and the same.

There are many other notions relevant here. For example, the ease in dealing with multi and nuisance parameters. Good references are available, however. See, for instance, Berger (1985).
1.3 Survey sampling: foundations

Consider a finite population \( \mathcal{N} = \{1, \ldots, N\} \) of \( N \) distinct and identifiable units, where \( N \) is known. Associated with each unit \( i \), there exists a real-valued characteristic \( y_i, i \in \mathcal{N} \). The vector \( y = (y_1, \ldots, y_N)^T \) is called the parameter vector with the parameter space \( \Theta \) being a subset of \( \mathbb{R}^N \). Our interest is to make inference about a parametric function \( \psi(\theta) \) with the population total \( \gamma = \sum_{i=1}^{N} y_i \) as a special case.

Let \( \mathcal{S} \) be the power set of \( \mathcal{N} \) and \( p(\mathcal{S}) \) be a probability measure defined on \( \mathcal{S} \). \( p(\mathcal{S}) \) is called a sampling design. If \( p(\mathcal{S}) \) does not depend of \( y \), then we say that \( p(\mathcal{S}) \) is a noninformative design. With the help of a sampling design \( p(\mathcal{S}) \), the survey practitioner chooses a sample \( \mathcal{s} \), subsequently he observes the corresponding \( y \)-values without any error. The sample \( \mathcal{s} \) together with the \( y \)-values, \( \{(i, y_i), i \in \mathcal{s}\} \), is called a data-point.

Now it is well known that, if we restrict to noninformative designs, then the statistic formed from \( \{(i, y_i), i \in \mathcal{s}\} \) by destroying the order and repetition of units is a minimal sufficient statistic; throughout the thesis, \( y \) stands for this sufficient statistic. (Also \( \mathcal{s} \) stands for the unobserved part of the label set, that is, \( \mathcal{s} = \mathcal{N} - \mathcal{s} \).) A general form of this result can be found in Sen and Sinha (1990).

Let \( p(\mathcal{S}) \) be a noninformative sampling design. Let \( y_s \) be a data-point observed using \( p(\mathcal{S}) \). If \( \Theta_s \) is the set of all parameter values which are consistent with \( y_s \), that is, \( \Theta_s = \{z \in \mathbb{R}^N : z_i = y_i \forall i \in \mathcal{s}\} \), then the likelihood function is given by

\[
L_p(y; y_s) = \begin{cases} 
p(\mathcal{S}) & \text{if } y \in \Theta_s \\
0 & \text{otherwise}
\end{cases}
\]

Thus the likelihood function tells us everything about the observed part of the population and nothing about the unobserved part. If \( p(\mathcal{S}) \) and \( q(\mathcal{S}) \) are two nonin-
formative sampling designs which resulted in the same data-point, then, for $y \in \Theta$, 

$$\frac{L_p(y; y_s)}{L_q(y; y_s)} = \frac{p(s)}{q(s)}$$

a constant. Thus, by the likelihood principle, inferential procedures should be free of the sampling design under consideration. The revolutionary implications of this result has already been mentioned in Section 1.1.

One subtle question now is, if one likes to stick to the likelihood principle, what is the role of randomization? The issue is unsettled till now; see Berger and Wolpert (1984) and Godambe (1982) for various views.

It is well known that Bayesian analysis adheres to the likelihood principle. Another point in support of Bayesian analysis in this context is the following: Suppose that the practitioner is interested in estimating $\gamma$. If a sampling design gave rise to the data point $y_s$, then,

$$\gamma = \sum_{i \in s} y_i + \sum_{i \notin s} y_i$$

with the first sum being fully known and the second sum being fully unknown. If $y_i$'s are not related in some way, how can we infer about the second sum by looking at the data point? Neyman found this link through the sampling distribution of a statistic, say $\sum_{i \in s} y_i$, where the variation is of the observed sample obtained by using the sampling design $p(.)$. Being based on hypothetical repetitions, it does not say anything about $\sum_{i \notin s} y_i$ from the data point $y_s$ at hand.

Now let us consider the Bayesian analysis for this problem. Suppose that $p(y)$ is a joint distribution for $y$ which expresses the practitioner's prior knowledge about the population parameter $y$. Then the posterior distribution of $y$ given the data point $y_s$ is given by

$$p(y | y_s) = \begin{cases} \frac{p(s)p(y)}{\int p(s)p(y)dy(s)} & \text{for } y \in \Theta, \\ 0 & \text{otherwise} \end{cases}$$
provided \( p(s) \) is a noninformative design. Observing that

\[
p(y(s) \mid y_s) = \frac{p(y)}{p(y_s)},
\]

we are actually considering the predictive distribution of the unobserved part \( y(s) = (y_i, i \in s)^T \) given the observed part \( y_s \). In this way, a predictive approach is inherent in the Bayesian analysis. This dual interpretation is utilized whenever it is appealing.

**Remark** Even if the sampling design \( p(s) \) depends on \( y_s \), but not on \( y(s) \), the conclusion that \( p(y \mid y_s) \) does not depend on \( p(s) \) still holds. For instance, sequential sampling design depend on \( y \) only through \( y_s \). A detailed discussion can be found in Zacks (1971). However, we will assume that sampling designs under consideration are noninformative.

### 1.4 Survey of literature

If \( p(y) \) summarizes the practitioner's prior information about the population characteristic \( y \), all the posterior inference can be made from \( p(y \mid y_s) \) or \( p(y(s) \mid y_s) \). For instance, if the problem is to estimate \( \gamma \), an appropriate transformation gives us the desired posterior distribution. If we assume a squared error loss function, the Bayes estimator for \( \gamma \) is the posterior expectation.

Thus the whole problem now lies in the choice of \( p(y) \). The literature on Bayesian analysis of survey sampling mainly concentrates on this aspect. Another important
problem is the choice of an optimal sampling design. Should the Bayesian argument be also used for a proper choice of sampling design? Perhaps yes! However not much work has been done in this direction. One exception is Zacks (1971). He proved Basu's (1971) and Godambe's (1966) claim that a Bayesian design for survey sampling is a purposive sequential design.

The simplest structure that one may like to put for \( p(y) \) is an exchangeable model. Basic results for such a model have been developed by Ericson (1969a) and Ericson (1970). Ericson (1969a) also investigated the relationship between an SRSWOR sampling design and an exchangeable population.

As exchangeability is too delicate to expect in reality, attempts have been made in search of other models. One simplest generalization of an exchangeable population is to consider a population where we can stratify in such a way that each subpopulation is exchangeable. This is in spirit of the classical theory of stratification. We refer to Ericson (1969b) for details.

While Ericson's models are parametric, his results related to exchangeable and stratified populations can be extended to nonparametric setup. Binder (1982) assumed a Dirichlet process prior for \( y \); Lo (1986, 1988) have also contributed to this area. In fact, Ericson's (1969a) model with Dirichlet-multinomial prior distribution is related to Dirichlet process analysis; Lo gives the details.

Scott and Smith (1969) gave a Bayesian analysis for two stage surveys with normal priors. Their approach can be treated as a conditional exchangeable model. These results are extended to multi-stage sampling by Malec and Sedransk (1983). A discrete version of Scott and Smith's work can be found in Nandram and Sedransk (1993).

Another model related to nonparametric Bayesian analysis is the Pólya posterior introduced by Meeden and Vardeman (1991). Pólya posteriors are closely related to the finite population bootstrap of Lo (1986).
Ericson's (1969a) seminal paper also considered populations with real auxiliary variables and its Bayesian analysis. Now many results are available in this direction. Bolfarine and others introduced many models; these can be found in Bolfarine and Zacks (1992). Some related results can be found in Särndal, Swensson and Wretman (1991) also.

While all these models are more or less attempt to justify classical procedures within Bayesian setup, a new area of research has been emerging; what we have in mind is the small area estimation. Bayesian analysis for this problem has many advantages. Malay Ghosh and others developed basic results in this area; in particular, we mention Datta and Ghosh (1992) and Ghosh and Lahiri (1990). We also like to mention some papers which propose empirical Bayes methods in small area estimation. The relevant papers are Ghosh and Lahiri (1987) and Ghosh and Meeden (1986) and a series of papers by Lahiri and R.C. Tiwari.

Other references are, nonresponse model by Basu and Pereira (1982) and time series model by Scott and Smith (1974). Some other references can be found in the thesis chapters.

While this survey is not intended to be complete, we feel that it gives a panoramic view of the existing results.

1.5 Summary of the thesis

This thesis has five chapters, including this introductory chapter. Contents of the remaining chapters are summarized in the following paragraphs.

In many contexts, besides having a study characteristic $y$, associated with each unit $i \in \mathcal{N}$ there is a known auxiliary measurement $x_i$. For estimating the population total $\gamma$, the survey practitioner can assist himself by utilizing these auxiliary
measurements, which are assumed to be a priori positively correlated with \( y \). When \((y_i / z_i)\)'s are approximately a constant in the population, Basu (1971) proposed an estimator for \( \gamma \) as

\[
e_B = \sum_{i \in s} y_i + \hat{\beta}_s \sum_{i \in s} z_i
\]

\[= \sum_{i \in s} y_i + \hat{\beta}_s (N \bar{z}_N - n(s) \bar{z}_s)\]

where \( \hat{\beta}_s = (1/n(s)) \sum_{i \in s} (y_i / z_i) \), \( \bar{z}_N = \sum_{i=1}^{N} z_i / N \), \( \bar{z}_s = \sum_{i \in s} z_i / n(s) \) and \( n(s) = | s | \).

Suppose the population z-mean \( \bar{z}_N \) is not known. However, it is much cheaper to obtain \( z \)-values as compared to \( y \)-values. Under this assumption we can use double sampling technique and modify the Basu's estimator as

\[
e_B = \sum_{i \in s(2)} y_i + \hat{\beta}_{s(2)} (N \bar{z}_{s(1)} - n(s(2)) \bar{z}_{s(2)})\]

where \( s(2) \subset s(1) \) and \( (s(1), s(2)) \) is obtained through a double sampling design.

We have studied various properties of \( e_B \), including admissibility and uniform admissibility. Details can be found in Chapter 2.

Another situation where double sampling method is commonly used in survey sampling is double sampling for stratification. In a stratified finite population it is usually assumed that the strata sizes are known a priori to the sample. But in practice this may not be the case. Double sampling for stratification is introduced in the classical theory to solve this problem. We give Bayesian analysis for this problem with the help of an analogy with a generalised animal abundance problem. The resulting posterior distribution of \( \gamma \) is rather complicated to evaluate analytically. We suggest the use of weighted bootstrap method to resolve this computational difficulty. These findings are useful in estimation of animal abundance also. Chapter 3 is devoted for this problem.
In survey sampling, problems often arise when the respondents are asked sensitive questions, for instance, about highly personal or controversial matter. To avoid providing the requested information, some respondents may intentionally give wrong answers. Thus, it is difficult to make inference about sensitive topics based upon survey sampling in which sensitive questions are asked directly.

Warner's randomized response model (Warner 1965) attempts to resolve this problem by injecting a probabilistic element into the questioning procedure. This method has got wide attention and many modifications and extensions are available now (cf. Chaudhuri and Mukherjee 1988). Winkler and Franklin (1979) has given a Bayesian analysis for this problem. However the posterior distribution of the population proportion is a rather complicated mixture distribution. In Chapter 4 we resolve this problem by using a simulation technique, called Gibbs sampler, introduced to statisticians by Gelfand and Smith (1990). We also give a Bayesian analysis for an empirical Bayes version of the Warner's model. We extent the techniques developed in the case of Warner's model to solve computational difficulties. These techniques are illustrated by a real data set reported by Liu and Chow (1976). A Bayesian analysis for the Simmons' unrelated question model is also presented in this chapter.

Estimation of the population total when the population has units with extremely large or extremely small characteristic values is an important and difficult problem. This is generally known as the problem of outliers. We give two useful methods of analysing this problem: (a) Bayes method and (b) empirical Bayes method. Here also, Bayes method utilizes the Gibbs sampler. While empirical Bayes method draws force from the inherent stratified nature of an outlier affected population. These methods are illustrated with a real data set reported in Ashok and Sukhatme (1976).

All remaining four chapters start with a set of introductory remarks. These
remarks are followed by a section, called preliminaries. In the preliminary section of each chapter we summarize important facts and tools used in that particular chapter. For instance, the preliminary section of Chapter 2 is devoted for the stepwise Bayes procedure.

There are two appendices for the thesis. In Appendix A, we present routine algebra and chapter-specific computer programmes. These are arranged according to chapters. In Appendix B, we present some random variate generation subroutines.

Throughout the thesis, we follow notations and conventions suggested by Lindley (1956). For instance, we do not distinguish between random variables and their observed values. However, to avoid ambiguity distinction is made whenever it is really needed.