BASU'S CONTRIBUTIONS TO THE FOUNDATIONS OF SAMPLE SURVEY

Glen Meeden, Department of Statistics, Iowa State University, Ames

Introduction

Whenever I read a paper by Dev I am impressed with the clarity of his writing and thinking. He is able to distill the essence of the topic at hand and present it in such a way that it seems almost obvious to me. This is particularly true in the foundations of sample survey where he has elegantly demonstrated the proper role of the sufficiency and likelihood principles. Because these principles fail to justify much of the current design based practice and because he has presented his arguments in a Bayesian context some survey samplers have chosen to either ignore or attempted to modify the consequence of these principles. This coldness to Bayesian ideas in survey sampling could be considered surprising since it is the one area in statistics where everyone agrees prior information should be used.

In the next section, the results of Basu and Ghosh (1967), which characterize the minimal sufficiency partition for discrete models, will be briefly summarized. In the third section, the results of Basu (1969) will be summarized. Here he demonstrated the role of the sufficiency and likelihood principles in sample survey, from which it follows, that once the sample has been drawn the inference should not depend in any way on the sampling design. In the fourth section, some of the implications of these results will be noted. In particular, the famous *Jumbo* example of Basu (1971) will be discussed. It will be shown how Basu's argument there suggests a pseudo-Bayesian approach to survey sampling. This approach is quite flexible in that one can incorporate various levels of prior information without specifying a prior distribution. Finally, the role of random sampling in survey sampling will be discussed briefly. It should be noted that Basu (1978) contains some further reflections on his earlier work.

Sufficiency in Discrete Models

For many years, in statistical decision theory, it has been an accepted convention, to begin by assuming the existence of a nonempty set X, equipped with a σ-algebra of subsets of X, say β, along with $\mathbb{P} = \{P_\theta | \theta \varepsilon \Omega\}$ a family of probability measures on (X, β). One of the consequences of Basu's work (along with others) was to fit survey sampling into this scheme. For such a model it is of interest to find the minimal sufficient statistic, assuming it exists. Now, in general, for such models a minimal sufficient statistic need not exist. However, for discrete models, which includes the sample survey model, a minimal sufficient statistic always exists and is easy to find.

The triple $(X, \beta, \mathbb{P})$ is said to be a discrete model if i) β is the class of all subsets of X and ii) each P_θ is a discrete probability measure. (We are also assuming that for each $x \varepsilon X$, there exists a $\theta \varepsilon \Omega$, such that $P_\theta(\{x\}) = P_\theta(x) > 0$.) Note that a discrete model is undominated if and only if X is uncountable.

Now a statistic is just a function, T, defined on X. By our choice of β every function T is measurable. Every statistic T defines an equivalence relation ($x \sim x'$ if $T(x) = T(x')$) on the space X. This leads to a partition of X into equivalent classes of points. Since we need not distinguish between statistics that induce the same partition of X, we may think of a statistic T as a partition $\{\pi\}$ of X into a family of mutually exclusive and collectively exhaustive parts π.

Using the usual measure theoretic definition of sufficiency one can prove the following factorization theorem for discrete models:

Theorem (Basu and Ghosh, 1967).

If $(X, \beta, \mathbb{P})$ is a discrete model, then a necessary and sufficient condition for a statistic (partition) $T = \{\pi\}$ to be sufficient is that there exists a real valued function g on X such that, for all $\theta \varepsilon \Omega$ and $x \varepsilon X$

$$P_\theta(x) = g(x) P_\theta(\pi_x)$$

where π_x is the part of the partition $\{\pi\}$ that contains x.

Using this theorem, it is easy to find the minimal sufficient partition for a discrete model. For each $x \varepsilon X$ let

$$\Omega_x = \{\theta |\ P_\theta(x) > 0\}.$$

Consider the binary relation on X: "$x \sim x'$ if $\Omega_x = \Omega_{x'}$, and $P_\theta(x)/P_\theta(x')$ is a constant in θ for all $\theta \varepsilon \Omega_x = \Omega_{x'}$." This is an equivalence relationship on X and defines the minimal sufficient partition.

The minimal sufficient statistic has an alternative characterization. For each $x \varepsilon X$ let $L_x(\theta)$ be the likelihood function, i.e.

$$\begin{aligned} L_x(\theta) &= P_x(\theta) && \text{for } \theta \varepsilon \Omega_x \\ &= 0 && \text{for } \theta \notin \Omega_x \end{aligned}$$

and

$$\bar{L}_x(\theta) = L_x(\theta) / \sup_\theta L_x(\theta)$$

be the standardized likelihood function. Consider the mapping

$$x \rightarrow \bar{L}_x(\cdot)$$

a mapping of X into a class of real-valued functions on Ω. This mapping is a minimal sufficient statistic, i.e. induces the minimal sufficient partition given above.

The Sufficiency and Likelihood Principles in Survey Sampling

The sufficiency and likelihood principles were widely used in other areas of statistics before their role in survey sampling was properly understood. The sufficiency principle states that if T is a sufficient statistic and $T(x) = T(x')$ then the inference about θ should be the same whether the sample is x or x'. This principle has gained wide acceptance. In discrete models since the mapping $x \rightarrow \bar{L}_x(\theta)$ is a minimal sufficient statistic, according to the sufficiency principle two sample points x and x' are equally informative if

$$\bar{L}_x(\theta) = \bar{L}_{x'}(\theta) \qquad \text{for all } \theta.$$

Note the sufficiency principle does not say anything about the nature of the information supplied by x. For this we need the likelihood principle which states that the information supplied by x is just the standardized likelihood function $\bar{L}_x(\theta)$.

To see the implications of these principles in survey sampling we consider a simple survey model. Let U denote a finite population of N units labeled 1, 2,...,N. Attached to unit i let y_i be the unknown value of some characteristic of interest. For this problem

$$\theta = (y_1,\ldots,y_N)$$

is the unknown state of nature. θ is assumed to belong to Ω a subset of N-dimensional Euclidean space, $\mathbb{R}^N$. The statistician usually has some prior information about $\underset{\sim}{y}$ and this could influence the choice of Ω. Often it is assumed that $\Omega = \mathbb{R}^N$ but this need not be so. We will assume that, in addition, associated with each unit i is m_i, a possible vector of other characteristics all of which are known to the statistician. We assume that the m_i's and their possible relationship to the y_i's summarize the statisticians prior information about $\underset{\sim}{y}$.

A subset s of $\{1, 2,\ldots,N\}$ is called a sample. Let $n(s)$ denote the number of elements belong to s. Let S denote the set of all possible samples. A (nonsequential) sampling design is a function Δ defined on S such that $\Delta(s)\varepsilon[0, 1]$ and $\sum_{s\varepsilon S}\Delta(s) = 1$. Given $\theta\varepsilon\Omega$ and $s = \{i_1,\ldots,i_{n(s)}\}$ where $1 \leq i_1 < \ldots < i_{n(s)} \leq N$ let $\theta(s) = (y_{i_1},\ldots,y_{i_{n(s)}})$. Suppose we wish to estimate the population total

$$\gamma(\theta) = \sum_{i=1}^{N} y_i$$

with squared error loss. Note $e(s, \theta)$ will denote an estimator of $\gamma(\theta)$ where

$e(s, \theta)$ depends on θ only through $\theta(s)$. If the design Δ is used in conjunction with the estimator e, then the risk function is

$$r(\theta; \Delta, e) = \sum_s \left[e(s, \theta) - \gamma(\theta)\right]^2 \Delta(s).$$

Typically a frequentist sampler uses the prior information summarized in the m_i's to choose some design Δ and then looks for estimators which are unbiased for estimating $\gamma(\theta)$. For such an unbiased estimator the risk function is just its variance.

For such a problem a typical sample point is the set of labels of the units contained in the observed sample along with their values of the characteristic of interest. We will denote such a point by

$$\begin{aligned} x &= (s, x_s) \\ &= \left(s, (x_{i_1}, \ldots, x_{i_{n(s)}})\right) \end{aligned}$$

when $s = \{i_1, \ldots, i_{n(s)}\}$ is the observed sample.

Hence for a given design Δ the sample space is given by

$$X = \{(s, x_s) | \Delta(s) > 0 \text{ and } x_s = \theta(s) \text{ for some } \theta \varepsilon \Omega\}.$$

So for a fixed $\theta \varepsilon \Omega$ the probability function over X is given by

$$\begin{aligned} P_\theta(x) = P_\theta(s, x_s) &= \Delta(s) \qquad \text{if } x_s = \theta(s) \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

This defines a discrete model. Note that

$$\begin{aligned} \Omega_x = \Omega_{(s,\, x_s)} &= \{\theta | P_\theta(x) > 0\} \\ &= \{\theta | \theta(s) = x_s\} \end{aligned}$$

from which it follows that

$$\begin{aligned} P_\theta(x) = P_\theta(s, x_s) &= \Delta(s) \qquad \text{if } \theta \varepsilon \Omega_x \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

If as before, $\bar{L}_x(\cdot)$ denotes the standardized likelihood function, we see that

$$\begin{aligned} \bar{L}_x(\theta) = \bar{L}_{(s,\ x_s)}(\theta) &= 1 \qquad \text{if } \theta \varepsilon \Omega_x \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

Since the mapping $x \rightarrow \bar{L}_x(\cdot)$ is a minimal sufficient statistic and the likelihood function is constant over Ω_x, all we learn from the observed data $x = (s,\ x_s)$ are the values of the characteristic for the units in sample and that the *true* θ must be consistent with these observed values.

Note that this observation is independent of the sampling design. That is, after the sample $x = (s,\ x_s)$ is observed the minimal sufficient statistic does not depend in any way on the value of $\Delta(s)$. (In fact, Basu demonstrated that this is true even for sequential sampling plans where the choice of a population unit at any stage is allowed to depend on the observed y-values of the previously selected units.) Furthermore, the principle of maximizing the likelihood function cannot be invoked to find an estimate of the population total since the standardized likelihood function is constant over Ω_x.

In the next section some implications of these results will be discussed.

Some Implications

For most statisticians, perhaps the most unsettling aspect of Basu's argument is his demonstration that the likelihood principle implies that the design probability should not be considered in analyzing the data, after the sample has been observed. In particular, choosing an estimator which is unbiased for a given design violates the likelihood principle. But from a naive point of view this is not surprising when one recalls the *strange* way probability is used in survey sampling. Since the characteristic y_i is assumed to be measured without error the only way probability enters the model is through the design Δ. That is the phenomenon of randomness is not inherent within the problem but is artificially injected into it by the statistician. In other areas of statistics the statistician uses probability theory to model uncontrollable randomness while in survey sampling the whole analysis is based on a controlled *randomness* introduced by the statistician.

Godambe (1966) had noted before Basu (1969) that the application of the likelihood principle to survey sampling would mean that the sampling design is irrelevant for data analysis. But he, as many other non-Bayesian statisticians since then, has chosen to ignore the likelihood principle and tried to justify a role for the design when analyzing the data.

Scott (1977) and Sugden and Smith (1984) considered situations where some information available to the person who designed the sample is not available to the one who must analyze the data. They argued that in such situations the design may become informative. Although such examples are interesting I do not feel that they lessen the force of Basu's argument.

Recall that the likelihood principle in survey sampling justifies a very intuitively appealing notion, that is, given the observed data $x = (s,\ x_s)$ one just learns the y_i's for $i \varepsilon s$ and that the unsampled y_j's for $j \not\varepsilon s$ must come from a θ which is consistent with x. So the basic question of survey sampling is how can one relate the unseen, $\theta(s') = \{y_j\colon\ j \not\varepsilon s\}$, to the seen, $\theta(s) = \{y_i\colon\ i \varepsilon s\}$. Without some assumptions about how these two sets are related, knowing $\theta(s)$ does not tell one anything at all about $\theta(s')$. Presumably, for a frequentist, the design Δ along with the unbiasedness requirement is a way to relate the unseen to the seen. But I have never understood the underlying logic of the relationship.

On the other hand, the Bayesian paradigm allows one to relate the unseen to the seen in a straightforward way which does not violate the likelihood principle. Let $q(\theta)$ denote the Bayesians' prior density over Ω. q would be chosen to represent and summarize the statisticians prior beliefs about θ. Given the sample $x = (s, x_s)$ one then computes the conditional density of θ given x, say $q(\theta|\ x)$. This is concentrated on the set Ω_x and is just q with the seen, $\theta(s) = \{y_i\colon\ i \varepsilon s\}$, inserted in their appropriate places and normalized, so it integrates to one over Ω_x. Then the Bayes estimator against q for the populational total is

$$\sum_{i \varepsilon s} y_i + \sum_{j \not\varepsilon s} E_q(y_j|\ x)$$

where for $j \not\varepsilon s$, $E_q(y_j|x)$ is the conditional expectation of y_j with respect to $q(\theta|x)$.

The form of the Bayes estimator emphasizes that estimation in survey sampling can be thought of as a prediction problem, i.e. of predicting the unseen from the seen. That is, in these problems one should argue conditionally from the seen to the unseen.

As was to be expected the Bayes estimator does not depend on the design. In most of the standard statistical decision problems an estimator is admissible if and only if it is a Bayes estimator or limit of Bayes estimators. This suggests that in survey sampling the admissibility of an estimator should not depend on a particular design. This was demonstrated in Scott (1975). Let Δ_1 and Δ_2 be two designs with Δ_1 dominating Δ_2, i.e. if s is such that $\Delta_1(s) = 0$ then $\Delta_2(s) = 0$ as well. Then Scott proved if the estimator e is admissible for design Δ_1 then it is also admissible for design Δ_2.

From the Bayesian point of view the statistician should use a design which minimizes the overall Bayes risk. In practice such designs are very difficult to find but often such minimizers are purposeful designs, i.e., designs which put probability one on a single set. Hence Basu has elegantly outlined a coherent theory of survey sampling in which random sampling or more generally the sampling design has little or no role to play. Ericson (1969) is one example of a Bayesian approach to survey sampling very much in the spirit of Basu. However, one serious difficulty in using a Bayesian approach to survey sampling is specifying a realistic prior distribution. Even for those who are somewhat sympathetic to Bayesian ideas, choosing a prior in survey sampling is almost impossible because of the larger number of parameters. Hence, it would be of interest to have an approach to survey sampling which did not violate the likelihood principle, allowed one to think conditionally given the sample, and allowed one to incorporate various levels of prior information relating the unseen to the seen without actually specifying a prior distribution. Such an approach is suggested in Basu's famous *Jumbo* example in Basu (1971).

Here Basu was discussing the Horvitz-Thompson estimator and other estimators which were suggested for some unequal probability designs. The Jumbo example dealt with estimating the total weight of a group of elephants where Jumbo was the largest.

Following Basu, let N be the size of the herd and y_i the weight of the i^{th} elephant. Let m_i be our best prior guess, before the sample is observed, of the weight of elephant i, that is, the m_i's incorporate all our prior information about the herd. We begin by assuming that the herd is reasonably homogeneous (in contrast to Basu, there is no Jumbo). Suppose a sample s with $n(s) = n > 1$ is chosen and the corresponding y_i's observed. Suppose we believe that these n observed ratios $\{y_i/m_i\colon i\varepsilon s\}$ are *representative* of the $N - n$ unobserved ratios $\{y_j/m_j\colon j\not\varepsilon s\}$. Although we may not be able to define *representative* we have an intuitive idea of what it means. Furthermore, if in practice we obtained a sample which we believed was not representative then we would be foolish to act as if it were.

Assuming the sample is representative then Basu suggested that $\bar{r} = \frac{1}{n} \sum\limits_{i\varepsilon s} (y_i/m_i)$ should be a good guess for y_j/m_j when $j\varepsilon s'$. Hence, for a typical unsample unit j, a reasonable estimate of y_j is $m_j\bar{r}$. This suggests a sensible estimate of the population total is

$$\sum_{i\varepsilon s} y_i + \left[\frac{1}{n} \sum_{i\varepsilon s} (y_i/m_i)\right] \sum_{j\not\varepsilon s} m_j. \tag{1}$$

This estimator can be given a pseudo Bayesian justification by creating a *posterior distribution* for the unseen given the seen which is appropriate when one believes the sample is representative. Suppose in the sample of n observations there are r distinct values of these ratios, say $\alpha_1,\ldots,\alpha_r$. Let k_j be the number of observed y_i/m_i's which are α_j for $j = 1,\ldots,r$. Construct an urn which contains n balls where k_j are labeled α_j for $j = 1,\ldots,r$. Then take as the pseudo posterior distribution for the $N - n$ unobserved ratios the distribution generated by simple Polya sampling from the urn. To begin, a ball is chosen at random from the urn and the observed value is given to the unobserved ratio with the smallest label. This ball and an additional ball with the same value are returned to the urn. Another ball is chosen from the urn and its value is given to the unobserved ratio with the next smallest label. This ball and another with the same value are returned to the urn. The process is continued until all $N - n$ unobserved ratios are given a value. We will call this pseudo posterior the *Polya posterior* for the unseen given the seen. The *Polya posterior* is a pseudo posterior because it does not arise from any single prior distribution over the parameter space. This is intuitively clear since it is data dependent. On the other hand, it does reflect the belief that the unseen are like the seen. Finally it is easy to check that the Bayes estimate of the population total using the *Polya posterior* is just the estimate given in (1).

Note in the special case when little is known about the herd, i.e. all the m_i's are equal, then the estimator in (1) reduces to $(N/n)\sum_{i\epsilon s} y_i$ which is the classical estimator of the population total.

In Meeden and Ghosh (1983) the estimator given in (1) was shown to be admissible. The proof used the stepwise Bayes technique. In the proof the *Polya posterior* played a crucial role. Hence, Basu's argument not only gives an intuitive justification for the estimator (1) but suggests a method for proving its admissibility. This approach can be extended to prove the admissibility of a variety of other estimators. (See Vardeman and Meeden (1984) for details.)

For example, suppose the population can be stratified into various strata each of which is relatively homogeneous. If the sample contains units from each stratum then the estimator in (1) can be used within each stratum, where within each stratum the m_i's are assumed to be equal, to produce an estimate of the population total. If in a given stratum, say k, we decide to sample n_k units then the stepwise Bayes argument shows that any set of n_k units within the stratum is optimal. That is, we may choose our n_k units by simple random sampling without replacement. This type of argument gives an noninformative Bayesian justification for a variety of the usual estimators in survey sampling along with a justification for choosing the sample at random.

One can argue that it is a relatively weak justification since it justifies any method of selecting the sample. In spite of Basu's arguments even some through going Bayesians, still admit to being attracted to the notion of randomization even though they do not know any intellectual justification for it. I however find Basu's statement on page 594 of Basu (1980), in slightly different context, quite compelling.

"I have no objection to prerandomization as such. Indeed, I think that the scientist ought to prerandomize and have the physical art of randomization properly witnessed and notarized. In this crooked world, how else can he avoid the charge of doctoring his own data?"

References

Basu, D. (1969): Role of sufficiency and likelihood principles in sample survey theory, *Sankhyā A* 31, 441-454.

Basu, D. (1971): An essay on the logical foundations of survey sampling, part one, *Foundations of Statistical Inference*, Holt, Reinhart and Winston, Toronto, 203-242.

Basu, D. (1978): *On the Relevance of Randomization in Data Analysis in Survey Sampling and Measurement*, N. K. Namboodiri, ed., Academic Press, New York, 267-292.

Basu, D. (1980): Randomization analysis of experimental data: The Fisher randomization test (with comments and rejoinder), *Journal of the American Statistical Association* 75, 575-595.

Basu, D. and Ghosh, J. K. (1967): Sufficient statistics in sampling from a finite universe, *Bull. Int. Stat. Inst.* 42, BK. 2, 850-859.

Godambe, V. P. (1966): A new approach to sampling from finite populations, I: Sufficiency and linear estimation, *Journal of the Royal Statistical Society B* 28, 310-319.

Meeden, G. and Ghosh, M. (1983): Choosing between experiments: Application to finite population sampling, *Annals of Statistics* 11, 296-305.

Scott, A. J. (1975): On admissibility and uniform admissibility in finite population sampling, *Annals of Statistics* 3, 489-491.

Scott, A. J. (1977): On the problem of randomization in survey sampling, *Sankhyā A* 39, 1-9.

Sugden, R. A. and Smith, T. M. F. (1984): Ignorable and informative designs in survey sampling inference, *Biometrika* 71, 495-506.

Vardeman, S. and Meeden, G. (1984): Admissible estimators for the total of a stratified population that employ prior information, *Annals of Statistics* 12, 675-684.