# Basu on Survey Sampling

**Glen Meeden**

Fifty years ago at a large scientific conference a statistician and a probabilist happen to set down together for lunch. In the ensuing small talk the probabilist admitted to knowing nothing about statistics and ask for a brief introduction to the subject. His companion outlined the common scenario of a company receiving a shipment of 1,000 widgets and selecting 20 of them at random to be tested. He then explained how the number of defective widgets in the sample could be used to make inferences about the state of the remaining 980 widgets in the shipment. The probabilist thought about this for a minute and then remarked, "I do not understand how knowledge about the 20 sampled units can tell me anything about the remaining 980 unsampled units." It is easy to forget how nonintuitive it is to claim that learning the observed values of the units in a sample, selected by random sampling, translates to knowledge about the unobserved values of the units remaining in the population.

If $y = (y_1, \ldots, y_N)$ is the vector of unknown population values of the characteristic of interest then given a sample $s$ we denote the observed or seen values by $y(s) = \{y_i : i \in s\}$ and the remaining unobserved or unseen values by $y(s') = \{y_j : j \notin s\}$. For Basu the fundamental question of survey sampling is how can one relate the seen to the unseen. Without some assumption about how these two sets are related knowing $y(s)$ does not tell one anything about $y(s')$. His application of the sufficiency and likelihood principles to survey sampling demonstrated that all we learn from the observed data are the values of the characteristic of interest in the sampled units and that the "true" vector of population values must be consistent with these observed values. Note this fact justifies the probabilist's statement. Moreover, Basu showed that this is true for any sampling plan where, at any stage, the choice of the next population unit to be observed is allowed to depend on the observed values of the characteristic of the previously selected units.

For Basu the Bayesian paradigm was the natural way to relate the unseen to the seen and still follow the likelihood principle. Let $\pi(y)$ be the prior density function or probability function for the Bayesian survey sampler over the parameter space of possible vectors $y$. The Bayesian selects $\pi(\cdot)$ to represent the prior information and his or her prior beliefs about $y$. Once the sample has been selected and the seen have been observed inferences are based on the posterior distribution, $\pi(y(s')|y(s))$, of the unseen given the seen and the design plays no role.

When Basu was writing we were, for the most part, restricted to prior distributions whose posterior distributions could only be studied using paper and pencil. With the recent advances in Bayesian computing it is now possible to simulate complete copies of $y(s')$ for many different possible posterior distributions. For such a posterior given $y(s)$ we can form many copies of $y(s')$ and hence many complete copies of the population. Supposes we are interested in estimating the function $\gamma(y)$. For

G. Meeden
Chairman and Head, School of Theoretical Statistics, University of Minnesota, Minneapolis, MN 55455
e-mail: glen@stat.umn.edu

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_6, 

each complete simulated copy we can compute the value of $\gamma$. Given a large set of such simulated values we can find approximately the corresponding point and interval estimates of $\gamma$. The key point in a Bayesian analysis is finding a sensible prior distribution. Once this in hand and the sample has been selected inferences can be made for any function $\gamma$ of interest.

This is in contrast to the design approach where the sampling design is often an important way to incorporate prior information into a problem. The design along with an unbiased requirement leads to an appropriate estimator. One difficulty with this approach is that each different choice of the function $\gamma$ requires a different argument. At a more fundamental level this suggests that the design approach does not yield a coherent method of relating the unseen to the seen. Basu never found this approach compelling because it violated the likelihood principle. Furthermore he never had much good to say about unequal probability sampling designs since, again by the likelihood principle, after the sample has been chosen the selection probabilities should play no role at the inferential stage.

Much of survey practice is still design based. It has always seemed curious to me that this one area of statistics where prior information is routinely employed makes use of this information in a way that cannot be justified from the Bayesian perspective. This is especially surprising given Basu's work. It is interesting to speculate why this is so. Part of the reason, I believe, is that it has always been difficult to find sensible and tractable prior distributions for large dimensional problems. This is particularly true in survey sampling which often deals with governmental statistics for which a certain degree of objectivity is expected. The challenge for a Bayesian is to find prior distributions which allow one to make use of the kinds of prior information which are now incorporated into a design. Our ability to now simulate complete copies of a population from more complicated but realistic posterior distributions should help fulfill the promise of Basu's work in the years ahead.