Basu on Survey Sampling

A.H. Welsh

"A circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants . . .", so begins Example 3 in Basu (1971), the most colorful and striking illustration of Basu's challenges to the design-based analysis of sample survey data. The full story is included in the box for easy reference. The owner decides to take a sample of size $n = 1$ ("As weighing an elephant is a cumbersome process") and is talked out of a non-random sample (select Sambo, the elephant that had the average weight 3 years before) and the model-based estimate $(50y)$ into an unequal probability sample (select Sambo with probability $99/100$ and any of the other elephants with probability $1/4900$) and the Horvitz-Thompson estimator ($100y/99$ if Sambo is selected and $4900y$ if any other elephant is selected). The point of the story is summarised in Figure 1 which shows the log-sampling distributions (i.e. the sampling distributions of the log of the estimators) for samples of size 1 of the model-based estimator and the Horvitz-Thompson estimator for a troupe of 50 elephants. (We plot the log-sampling distributions to improve the visual impact.) On this scale, the model-based estimator is very close to the actual total weight (indicated by an arrow) but, and this is Basu's elegantly made point, the design-unbiased Horvitz-Thompson is far from the actual total weight in every possible sample. The design-based optimality of the Horvitz-Thompson estimator is no consolation to either the circus owner or the "unhappy statistician" who, Basu tells us, "lost his circus job (and perhaps became a teacher of statistics!)".

The elephant story provokes and challenges, delights and frustrates, and ultimately encourages deep thinking on serious issues. Basu argued that the analysis of survey data should be subject to the same general principles as the analysis of other forms of data, and that there should be no special pleading for survey analysis to be treated differently from other statistical analyses. It is not surprising therefore that Basu's elephants illustrate specific points about survey analysis as well as general points about statistical analysis. In the survey context, Basu's elephants illustrate specific difficulties with unbiased estimation, unequal probability samples and design-based analysis. The elephant's bring these together with striking effect but they can also be teased apart and considered separately. One response to the example (Hajek in the discussion) is to suggest a different estimator for θ, the total weight of the elephants: since we know the weights of the elephants from the last time they were weighed, we should use a ratio estimator rather than the Horvitz-Thompson estimator. It is a nice irony that the ratio estimator is slightly design-biased! For another suggestion, see Rao in the discussion of Basu (1978). A different response is to suggest that we use a different design, perhaps with less variable weights, the extreme choice being equal probability sampling with all the weights equal. The

A.H. Welsh (✉)
Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia
e-mail: alan.welsh@anu.edu.au

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_11,

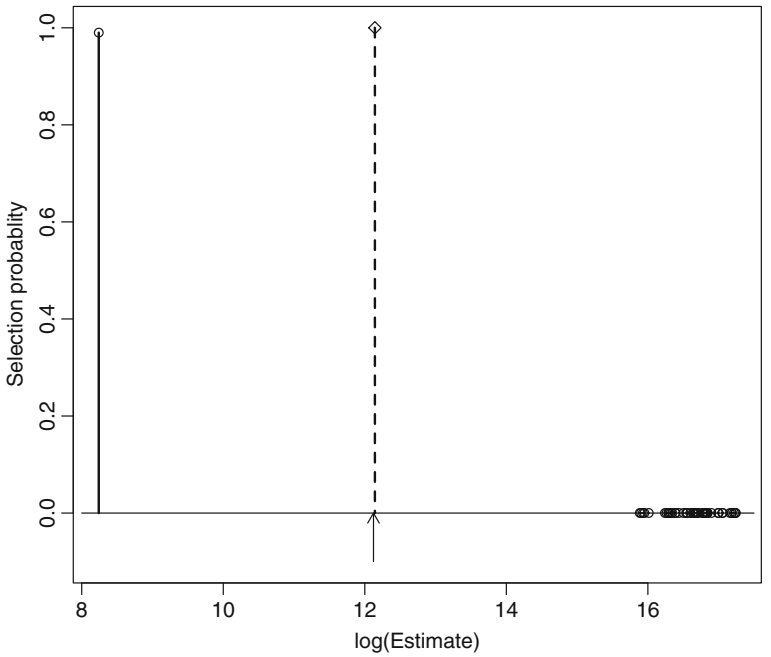

Fig. 1 The design-based log-sampling distributions (i.e. the x-axis is on the log scale) of the model-based and Horvitz-Thompson estimators for a sample of size 1 from a troupe of 50 elephants. The model-based estimator has a degenerate distribution represented by the diamond symbol and the dashed segment. The Horvitz-Thompson estimator has a distribution represented by the circles and solid segments. The true total weight of the troupe is shown by the arrow below the $y = 0$ line

elephants show that if we use unequal probability sampling and the weights do not depend on θ but simply reflect our desire or otherwise to include each unit in the sample, giving the most weight to the observations we want to include least in the sample may not be sensible. Actually, as pointed out in the discussion by Hajek, Godambe and Koop, the sampling design may incorporate prior information about the population and hence depend on θ, but a relationship like this is very difficult to formalize mathematically and so difficult to exploit. Whether for this or for some other reason, Basu did not see value in unequal probability sampling, even though, in simple examples, he did explore some purposive designs for which the selection probabilities are highly unequal. Basu's preferred response, and the motivation for the example, is for us to do a different kind of (non-design-based) analysis which does not depend on the sampling design.

Basu's critique is much broader than unbiased estimation and unequal probability sampling: The fundamental point in his (later) sample survey papers is that the design-based approach contravenes the likelihood principle and hence should not be used for the analysis of survey data. One could argue this from the point of view that there is no likelihood in the design-based framework, although this would open the possible rejoinder that the likelihood principle is then not relevant. Instead, Basu argued that there is a likelihood, the function that equals the probability of selecting the given sample on Ω_x, "the set of parameter points that are consistent with a given sample", and zero otherwise (Basu, 1969). If the sampling design does not depend on the target parameter θ, the design is ancillary and the likelihood is constant on Ω_x. If the ith elephant weighs Y_i, then $\theta = \sum_{i=1}^{N} Y_i$ and, if we sample a single elephant weighing y, the likelihood is constant on the set $\Omega_x = \{\theta \geq y\}$. It is interesting that this likelihood is derived from the sampling design and seems to require a probability sample: If the sample is purposive, there is nothing stochastic in the setup so, although we can simply define the likelihood to be constant on Ω_x and zero elsewhere, this function is not the joint density of the sample

viewed as function of the unknown parameter. Basu felt that we should implement the likelihood principle by doing a Bayesian analysis so he would have specified a prior on θ, thereby introducing a stochastic element, but this still leaves open the question of how to interpret the likelihood which is defined without reference to the prior.

The model-based approach provides another way of introducing probability into survey analysis (by treating $\omega = (Y_1, \ldots, Y_n)$ as a random vector) and hence of obtaining a likelihood (from the distribution of ω). How does this relate to what Basu (1978) called his "neo-Bayesian thesis on sample surveys"? Royall raised the model-based approach in his discussion of Basu (1971) but Basu, although he may have intended to, did not really engage with it. He suggested in response to Royall that superpopulation models are exactly like a Bayesian formulation of the background knowledge. He argued in his (1978) response to the discussion that superpopulation models are not objective and do not even exist, except in the mind, something he presumably also felt of his prior distributions. Since a prior for θ implies a distribution on ω from which the prior can be derived, Basu's neo-Bayesian and the model-based analysis should be able to be put into close numerical agreement by making compatible choices of prior and superpopulation model. These choices are potentially checkable from a census, at least in some cases and at least to the same extent as ordinary statistical models. To this extent at least, they do have an objective existence. It is interesting to explore these issues in more detail in the context of a different Basu example.

In Example 1 from Basu (1978), the population consists of N units, each of which is either defective ($Y_i = 1$) or non-defective ($Y_i = 0$). The units are produced by a mechanical device in such a way that, after the first defective unit, all the rest of the units are defective. The problem is to estimate the number of defective units $\theta = \sum_{i=1}^{N} Y_i$ from the values Y_i observed on a sample s of units. Let v be the largest $i \in s$ such that $Y_i = 0$; if $Y_i = 1$ for all $i \in s$, set $v = 0$. Let w be the smallest $i \in s$ such that $Y_i = 1$; if $Y_i = 0$ for all $i \in s$, set $w = N + 1$. Basu pointed out that, with probability one, $\theta \in \Omega_x = [N - w + 1, N - v]$ and this implies that some samples are much more informative than others: the best sample has $w = v + 1$ because then we know θ exactly. Although it is not our primary concern here, the design-based analysis of the example is interesting too because it explicitly permits us, if we so choose, to ignore the structure of the population. For example, if we select the sample by simple random sampling without replacement, we can use the expansion estimator $\hat{\theta}_E = (N/n) \sum_{i \in s} Y_i$ to produce an optimal, design-unbiased estimator of θ which is often outside the interval Ω_x. Of course, we expect to do better by incorporating the structure of the population into the sampling design. For example, we can select the first unit at random; at the $(k + 1)$th step, select the next unit at random in $[v^{(k)} + 1, w^{(k)} - 1]$, where $v^{(k)}$ and $w^{(k)}$ are the values of v and v from the first k observations; and continue until $w^{(k)} = v^{(k)} + 1$ or $k = n$. This is a stochastic version of the purposive design discussed by Basu which might be used in a design-based analysis, provided the design-based analyst can work out the sample inclusion probabilities needed to construct an estimator of θ. However, Basu would still have criticised this analysis as being in conflict with the likelihood principle.

Basu did not present his own analysis for this example, but we can construct an analysis he might have agreed to. For a sampling design which does not depend on θ, the likelihood is constant on the interval $[N - w + 1, N - v]$ so, if the prior density of θ is $q(\theta)$, the posterior density is

$$q(\theta|x) = \frac{q(\theta)}{\sum_{t=N-w+1}^{N-v} q(t)}, \qquad \theta = N - w + 1, N - w + 2, \ldots, N - v,$$

(Basu, 1969). When we are interested in a point estimate of θ, we use the posterior mean

$$\hat{\theta}_q = \frac{\sum_{t=N-w+1}^{N-v} t q(t)}{\sum_{t=N-w+1}^{N-v} q(t)}.$$

In the model-based approach, we model the distribution of ω. For this particular population, it is completely equivalent to model θ or the label of the first defective unit $M = N - \theta + 1$. The optimal mean squared error predictor of θ is given by $\hat{\theta}_p = N + 1 - E(M|s) = N - w + 1 + E(w - M|s)$, where the second expression is written in the familiar form of a sample contribution plus a non-sample contribution. Now we know that $Y_i = 0$ for $i \leq v$ and $Y_i = 1$ for $i \geq w$ so the sample information is that $m \in [v + 1, w]$. It follows that

$$P(M = m|s) = P(M = m|v + 1 \leq M \leq w) = \frac{P(M = m)}{\sum_{k=v+1}^{w} P(M = k)} \quad m = v + 1, v + 2, \ldots, w,$$

from which we can compute $E(M|s)$ and hence

$$\hat{\theta}_p = N + 1 - \frac{\sum_{k=v+1}^{w} kP(M = k)}{\sum_{k=v+1}^{w} P(M = k)}.$$

Algebraically, $\hat{\theta}_p$ equals the posterior mean $\hat{\theta}_q$ when $q(t) = P(M = N + 1 - t)$ (i.e. q is the distribution of $\theta = N + 1 - M$), supporting Basu's (1971) response to Royall. Adopting a prior for θ is implicitly adopting a distribution for ω and, at least numerically, the consequences can be made to match. For example, if we model M as having a uniform distribution on $\{0, 1, \ldots, N\}$, then

$$\hat{\theta}_p = N - w + 1 + (w - v)/2$$

and this equals the posterior mean $\hat{\theta}_q$ when the prior for θ is uniform on $\{0, 1, \ldots, N\}$. If instead we model M as having a geometric distribution with parameter π, then we can show that

$$\hat{\theta}_p = N - w + 1 + \left[w - \frac{1 + v\pi - (1 - \pi)^{w-v}(1 + w\pi)}{\pi\{1 - (1 - \pi)^{w-v}\}} \right].$$

and this equals the posterior mean $\hat{\theta}_q$ when the prior for θ is the distribution of $\theta = N + 1 - M$ and M has a geometric distribution with parameter π. Importantly, algebraic equality does not mean that $\hat{\theta}_q$ and $\hat{\theta}_p$ have the same content and meaning. From Basu's Bayesian perspective, π is a hyperparameter which we are free to specify: as $\pi \to 0$, $\hat{\theta}_q \to N - w + 1 + (w - v - 1)/2$ and as $\pi \to 1$, $\hat{\theta}_q \to N - v$ so that $\hat{\theta}_q \in \Omega_x \backslash [N - w + 1, N - w + 1 + (w - v - 1)/2]$. The data tells us that $\theta \in \Omega_x$ and the prior selects a particular point in Ω_x to resolve the arbitrariness of where, but the choice is entirely driven by us through the prior rather than the data. (If the prior has no support on Ω_x, the sample represents an event with prior probability zero and there is no usable posterior distribution. Thus, even if we have strong beliefs about the value of θ, the prior should still put some probability on every possible value of θ.) From the model-based perspective, π is an unknown parameter which we need to estimate. An often attractive way to do this is to use the maximum (model-based) likelihood estimator. The model-based likelihood is obtained from the population density of M (in this case, the geometric distribution) by treating ω as the complete data, the sample as incomplete data with the non-sample data missing (in a way determined by the design) and summing the complete data likelihood over the unobserved data. If we assume the sampling design is uninformative (i.e. sample selection does not depend on ω), then we can proceed straightforwardly. The model-based likelihood is not the same as the design-based likelihood used by Basu. Nonetheless it is also a likelihood, so we can impose a prior distribution on π and then do a Bayesian analysis which can be viewed as a hierarchical version of Basu's analysis with a hyperprior on π. Alternatively, we can substitute $\hat{\pi}$ into $\hat{\theta}_q$ as we do in the model-based analysis and view the result as an empirical Bayes predictor. Basu only considered enumerative inference about finite-population parameters like θ and had no interest

in analytic inference about hyperparameters like π; the model-based approach allows us to make analytic inferences about parameters like π as well as enumerative inference about finite-population parameters like θ. If we were to pursue the analysis beyond point estimation to inference, Basu would have objected to our using the sampling distribution in the model-based analysis and insisted on using his posterior distribution.

It is a challenge to describe how Basu would have analysed specific examples because he wrote more about what he would not do than what he would do and, when asked specifically, declined to provide more than general comments. In his very brief response to the discussion to Basu (1978), he wrote that analysing survey (probably meaning any) data is more an art than a science and he could say no more than that the analysis should be Bayesian (in the sense of fixing the sample and speculating about the parameters). Basu (1978) was clear that we need to know how the data were collected in order to analyse them - but, other than explicitly rejecting the design-based approach to doing this, he did not explain how to incorporate the data collection process into the analysis. It is natural for a Bayesian to include it in the prior specification, although this may be very difficult to achieve, particularly with purposive sampling. One possible role for probability sampling then is to simplify the way the data were collected and hence the prior specification. Basu (1969, 1978) also argued that except in simple populations, purposive sampling is too hard to justify (although, as Rao pointed out in the discussion to Basu (1978), this is not the case with Royall's purposive designs) and probability sampling can help a statistician defend his or her integrity. Basu's views on the role of randomisation are close to those of Royall (1976) and Rubin (1978).

Basu's papers on survey sampling should be read by everyone with an interest in survey sampling, indeed in statistics. The discussion papers are the most stimulating: they can be read starting with Basu (1971), referring to Basu (1969) for technical support, and then Basu (1978) or the other way round. The discussions and responses from Basu enhance the papers, stimulating much further thought. The paper by Basu and Ghosh (1967) on sufficiency is written in a very different, much more technical style. Basu (1958) is a traditional design-based paper, written before Basu became a Bayesian. It does not challenge the basic design-based framework in the same way as the later papers but, and this is characteristic of Basu and one of the reason his papers are still so valuable, it does challenge the usual method of analysing samples collected with replacement. With hindsight, it is tempting to see hints in Basu (1958) of what was to come, but it is a long way from there to the elephants, the circus and the "unhappy statistician". Statistics has benefitted enormously from the fact that Basu made that journey, questioning each step of the way.

References

Basu, D. (1958). Sampling with and without replacements. *Sankhya* **20**, 287–294.

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhya* **31**, 441–454.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part I (with discussion). In *Foundations of Statistical Inference*, eds V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart and Winston, 203–243.

Basu, D. (1978). On the relevance of randomization in data analysis (with discussion). In *Survey Sampling and Measurement*, ed N.K. Namboodiri, New York: Academic Press, 267–339.

Basu, D. and Ghosh, J.K. (1967). Sufficient statistics in sampling from a finite universe. *Bull. Int. Statist. Inst.* **42**, 850–859.

Royall, R.M. (1976). Current advances in sampling theory: Implications for human observational studies (with discussion). *Amer. J. Epidem.* **104** 463–477.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomisation. *Ann. Statist.*, **6** 34–58.