SURVEY SAMPLING – AS I UNDERSTAND IT
(A Development of Optimality Criterion)

V. P. Godambe, University of Waterloo

This was the Gold Medalist Presentation at the Statistical Society of Canada meetings held in Victoria, 6th June 1988.

> For since the fabric of the universe is most perfect and the work of a most wise Creator, nothing at all takes place in the universe in which some rule of maximum or minimum does not appear.
> – Leonhard Euler

Introduction

This is a brief overview of the historical development of the optimality criterion in survey sampling theory and practice. The presentation here has been considerably simplified for it takes for granted a fundamental result. In survey sampling set-up the entire data can be effectively summarized by the set of observed units (or individual labels) together with the corresponding variate values as in (1) to follow. This is a basic discovery due to Basu. He (1958) proved that in survey-sampling set-up (1) constitutes a *minimal sufficient statistic.*

Definitions, Notation and the Problem

Survey Population P is a finite collection of individuals (houses, blocks, farms, households, etc.), each bearing a distinctive label i; we may write

$$P = \{i{:}1,\ldots,N\},$$

where N is the size of P. Variate under study such as income, size, produce, etc. is denoted by y. The value of y associated with the individual i is y_i, $i = 1,\ldots,N$.

We want to estimate some *unknown* characteristic, say the mean

$$\bar{Y} = \sum_1^N y_i \Big/ N$$

of the population P. For this purpose a sample s of size n is drawn from P ($s \subset P$), using a *sampling design* (simple random sampling or stratified sampling, etc.) and the values y_i, $i \in s$ are ascertained through a survey.

Problem I: To estimate $\bar{Y}$ given the data

$$d = \{(i, y_i) : i \in s\} \tag{1}$$

and the sampling design. (A related problem is, how to use the pre-survey knowledge about P, particularly in the choice of a sampling design?)

For historical reasons the above problem remained confused, until recently, with the following quite different problem.

A treatment is tried n times with the following results

$$y_1, y_2, \ldots, y_n. \tag{2}$$

Problem II: To estimate the average treatment effect θ on the basis of the data (2).

Fundamental Distinction

The fundamental distinction between the two problems above becomes at once clear by the fact that while in problem (II), the sample mean $\sum_1^n y_i/n$ is the *unbiased minimum variance* (UMV) estimate for "θ", the corresponding mean $\sum_{i \in s} y_i/n$ in problem (I) is *not* UMV for $\bar{Y}$, even for a simple random sampling design.

The above phenomenon, as is now well understood, is due to the existence of individual labels "i" in the data (1), unlike in data (2). "$\bar{Y}$" in problem (I) is the mean of the *actual* (survey) population. In contrast "θ" in problem (II), is the mean of a *hypothetical* population generated by repeated (independent) trials of the treatment.

Why was problem (I) confused with problem (II) for a long time?

Answer: When the survey sampler arrived on the statistical stage (at about the beginning of this century), there already was a statistical theory developed by Galton, Pearson and others (to study primarily biological phenomena) which essentially dealt with problems akin to (II) of *hypothetical* populations. The confusion arose out of the attempts of the early survey samplers to use the then existing statistical theory to solve problem (I) concerning the *actual* (survey) population.

Historical Comments

Today's popular understanding of statistics consists of probabilistic estimates, say for instance, of country's average income, based on some random samples. But essentially this meshing of probability calculus with actual social statistics, historically proved to be far more formidable than establishing central limit theorem or Bayes theorem and the like. Actually both social statistics (Graunt) and probability theory (Pascal & Fermat) originated around 1660, but the meshing of the two occurred only in this century. Even in earlier history (for instance Jewish & Jain literature) one can find discussions of *uncertain* (probabilistic) inference; almost none relate to survey sampling. One exception I have temptation to quote. This is from Mahabharat, the old Indian epic (Vana-Parva; Nala-Damayanti Akhyan).

The God Kali has his eye on a beautiful princess and is dismayed when Nala wins her hand. In revenge an evil spirit enters the body of the virtuous prince. Crazed with frenzy for gambling, Nala loses his kingdom, and wanders demented for many years. Nala's change of fortune is described in a remarkable anecdote.

In an alien form, he has been travelling with another king, Bhangasuri. This latter, wanting to flaunt his skill in numbers, estimates the number of leaves, and the number of fruit, on two great branches of a spreading tree. There are, he avers, 2,095 fruits. Nala counts all night and is duly amazed by morning. Bhangasuri accepts his due:

I of dice possess the science, and in numbers thus am skilled.

He agrees to teach this science to Nala in exchange for some classes in horsemanship, in which, despite his exile, Nala still excels. At the end of this sensational course in survey-sampling Nala vomits out the poison of Kali, and is restored his normal form. Kali, exorcised by mathematics, retires to the tree. Nala returns to his kingdom, offers his still faithful bride as his final stake and quickly recoups all his losses, and lives happily ever after.

(Reproduced from History and Philosophy of Science Seminar by Ian Hacking)

Neyman's UMV-Criterion

The first well publicized attempt to solve the survey sampling problem, Problem I, using the then available statistical theory developed by Galton, Pearson, Fisher and others was due to Neyman, 1934. Actually this theory, as said before, was meant for hypothetical populations of Problem II. Following this theory, for simple random sampling (with or without replacement) Neyman considered the class of unbiased estimates (for the population mean $\bar{Y}$) of the form

$$\sum_{r=1}^{n} a_r y_r$$

where a_r is the coefficient associated with the r^{th} draw and y_r is the observed value of y at the r^{th} draw. The variance of this estimate is minimized, Neyman argued, using Gauss-Markov theorem, for $a_r = 1/n$, $r = 1,\ldots,n$. In this sense, Neyman demonstrated the UMV-ness of the sample mean. Similarly for stratified sampling he established UMV-ness of the corresponding weighted mean. (Similar previous, but little known results are due to Tchuprow (1923); see

Bellhouse, 1987.) In retrospect it appears Neyman obtained UMV estimates by restricting himself to the class of estimates which depended on individual labels i, *only* to the extent they determined the stratum to which the individual belonged. That is, *labels were ignored within each stratum.*

For several years, following Neyman, survey samplers investigated UMV estimation for more sophisticated designs than stratification. For reducing variance of estimates Hansen and Hurwitz (1943) introduced unequal probability sampling. Here however individual labels (i) were used not just for stratification but also were used even *within strata.* That is, in a stratum, two individuals could be selected with different probabilities.

What happened to Neyman's UMV-estimation here? Using individual labels i, Horwitz and Thompson (1952) constructed three different classes of estimates and investigated UMV estimation in each class. Though these latter investigations were inconclusive, the work clearly established that wider classes of estimates, than those considered by Neyman, could be constructed, using individual labels.

Neyman's introduction of UMV estimation in survey sampling led to an improved practice of stratified sampling, a better understanding of randomization and finally suggested the innovation of unequal probability sampling and general sampling designs.

Here however, the UMV-criterion appeared to have reached its limits of usefulness.

During 1935-1955 and even afterwards, while comparing variances of different estimates, possibly under different designs, proved to be rewarding, a search for UMV estimation led to futile confusion mentioned earlier; for such estimation was generally nonexistent!

Godambe (1955) introduced a general class of label dependent estimates of which all the known estimates were special cases. For this class, he demonstrated that UMV estimation was nonexistent, for any sampling designs (trivial exceptions apart). Particularly the sample mean was *not* UMV for the simple random sampling design.

Looking back, it would appear that survey samplers made considerable progress in sampling practice and theory, in their search for the nonexistent UMV estimation! But such things can happen in Science. Or one may say, survey samplers, in their investigation of UMV estimation, *informally* restricted themselves to the use of labels only to the extent they *intuitively* looked useful. This was the case with Neyman (1934). For a general development of this approach we refer to Hartley and Rao (1968).

A New Criterion: UMSV

Godambe (1955) also showed that in the class of all (linear) label dependent unbiased estimates, for the population mean $\bar{Y}$, the HT-estimate

$$e_{HT} = \frac{1}{N} \sum_{i \in s} y_i/\pi_i, \tag{3}$$

(due to Horwitz and Thompson, 1952), where π_i is the *probability of including the individual* i in the sample s drawn by the specified sampling design, has minimum *expected* variance. Here expectation is w.r.t. any distribution belonging to a *class* of distributions on the variate values $(y_1,\ldots,y_i,\ldots,y_N)$ under study. This class of distributions, called a *Superpopulation Model* (SPM), is supposed to be a formalization of our pre-survey knowledge of the survey-population P (see next section for illustration). Thus w.r.t. the SPM the HT-estimate is UM$\mathcal{E}$V: $U \equiv$ unbiased, $V \equiv$ variance, w.r.t. sampling design and $\mathcal{E} \equiv$ expected w.r.t. the SPM. Note that many estimates in common use, such as the sample mean for simple random sampling and the appropriately weighted mean for stratified sampling are but special cases of the estimate e_{HT} in (3). Hence they are UM$\mathcal{E}$V w.r.t. suitable SPMs.

Actually, since much earlier than 1955, variances had been compared in terms of their expectations w.r.t. the SPM (Cochran, 1939). Thus in the absence of UMV-estimation its replacement by UM$\mathcal{E}$V-estimation seemed natural. By now UM$\mathcal{E}$V-criterion seems to have received a general acceptance in theory as well as in practice. It is also used, somewhat reluctantly though, by *Model Theorists* in sampling.

The discovery that in survey-sampling the likelihood function is independent of the sampling design and hence according to the "*Likelihood Principle*" (LP) the inference must be independent of the design (randomization) probabilities (Godambe, 1966), gave impetus to the development of the model theory (Royall, 1970). This theory, to implement the above conclusion of LP, *restricts* inference/estimation exclusively to the probabilities given by SPM. (Such restriction was previously proposed by Brewer (1963), but he did not tie it to the LP. For this reason, possibly, Brewer's work was not effective in the development of model theory. By this time due to the works of Barnard, Birnbaum and Savage, LP became respectable.) With this restriction, the model theory estimation, using the notation above, proceeds as follows.

For a *given (fixed)* sample s, in terms of y_i: $i \in s$ construct the *class* of all linear estimates which are SPM-unbiased for the survey-population characteristic, say $\bar{Y}$. From this class, the minimum variance estimate (SPM-UMV) is recommended, for practical use, by the model theory.

Now for any sample s, the SPM-UMV estimate exists for rather restrictive SPMs. On the other hand when design and model probabilities are combined one can obtain UM$\mathcal{E}$V estimates (or close approximations) for far more flexible SPMs incorporating nuisance parameters of high dimension (Godambe, 1982, 1983).

Anyway even model theorists, in an attempt to make their estimation robust (to departures from the assumed SPM), have relied on the UM$\mathcal{E}$V-criterion (Brewer, 1979). Actually, from the model theorists ideological criticism and rejection, randomization emerged with new meaning, vigor and applications.

As I mentioned before, the UMV-criterion led to better understanding and practice of stratified sampling; the same thing can be said to have been

achieved by the UMŠV-criterion for unequal probability sampling beyond stratification.

Yet, the UMŠV-criterion is rather restrictive. It is generally non-vacuous only for fixed sample size designs. As mentioned before the HT-estimator is UMŠV—but generally only for fixed sample size designs. It is absurd for the following (rather *extreme*) random sample size design: with probability 1/2, a random sample of size "1" is drawn, and with remaining probability 1/2, the *whole population* is sampled. Now when the whole population is sampled, the *HT*-estimate (3) of the population mean $\bar{Y}$ is approximately $2\bar{Y}$!. Yet *random sample size* designs do occur in practice. For instance in surveys having *non-respondents* the (effective) sample size is essentially a random variate. The same thing happens for domain estimation.

Just as the extension of the UMV criterion to the UMŠV criterion was necessary to cover label dependent estimates, a further *extension* of the UMŠV itself is necessary to cover *random sample size designs.* This is achieved by the UMŠV-f criterion introduced in the next section. With this introduction, we can use even more flexible/broader SPMs than was possible under the UMŠV criterion. This will be clear soon.

UMŠV-f Criterion

Here we present the work of Godambe and Thompson (1986a). In addition to the notation above we denote by x_i the *covariate* value associated with the individual i, $i = 1,\ldots,N$. We *assume* $\boldsymbol{x} = (x_1,\ldots,x_i,\ldots,x_N)$ *known* and the SPM to be a *class* of distributions on $(y_1,\ldots,y_N)$ satisfying the following conditions:

(I) Given the covariate $\boldsymbol{x}$, $\boldsymbol{y} = (y_1,\ldots,y_i,\ldots,y_N)$ are distributed mutually independently.

(II) With respect to any distribution in the class the expectation $\mathcal{E}(y_i - \theta x_i) = 0$, $i = 1,\ldots,N$.

That is, under the SPM, θ is the regression parameter, intercept terms being ignored for simplicity. We define

$$\tilde{g} = \sum_1^N (y_i - \theta x_i); \tag{4}$$

$\tilde{g}$ is said to be a population or $\boldsymbol{y}$-based *unbiased estimating function*, since $\mathcal{E}(\tilde{g}) = 0$. If $[\tilde{g} = 0] \Rightarrow [\theta = \theta_N]$, θ_N is a $\boldsymbol{y}$-based estimate of the SPM-parameter θ. Further $\theta_N = \left(\sum_1^N y_i / \sum_1^N x_i\right)$ is itself a Survey Population parameter. Godambe and Thompson (1986a) theory provides *optimal* (sample based) estimation for θ_N as follows: Let $h(d, \theta)$ be any function of the parameter θ and the data $d = \{(i, y_i) : i \in s\}$ in (1), with

$$E(h - \tilde{g}) = 0, \tag{5}$$

"E" being the expectation under the sampling design, holding $\boldsymbol{y}$ and $\tilde{g}$ in (I) and (4) above fixed. The function h satisfying (5) is called a (design) *unbiased* estimating function; a solution of the equation $h(d, \theta) = 0$, provides an (data d based) estimate of both the parameters θ_N and θ. Now the function $h^*(d, \theta)$ satisfying (5) is said to be UMSV-f Optimum (f for estimating function), if for any h satisfying (5)

$$\mathcal{S}E(h^* - \tilde{g})^2 \leq \mathcal{S}E(h - \tilde{g})^2 \tag{6}$$

where $\mathcal{S}$ as before is the expectation w.r.t. the SPM-I&II above.

Theorem. For SPM-I&II, and any sampling design with $\pi_i > 0$, $i = 1,\ldots,N$, UMSV-f h^* is given by

$$h^* = \sum_{i \in s} (y_i - \theta x_i)/\pi_i. \tag{7}$$

Solving the equation $h^* = 0$, we get for θ and θ_N, the optimum estimate

$$e = \frac{\sum_{i \in s}(y_i/\pi_i)}{\sum_{i \in s}(x_i/\pi_i)}. \tag{8}$$

As a special case for all $x_i \equiv 1$, in (8),

$$e = \frac{\sum_{i \in s}(y_i/\pi_i)}{\sum_{i \in s}(1/\pi_i)}. \tag{9}$$

The relationship between the estimates e in (9) and the HT-estimate e_{HT} in (3) is given by the fact that for any sampling design

$$E\left\{\sum_{i \in s}(1/\pi_i)\right\} = N.$$

Note, now, for the random sample size design, considered before in previous section, $\pi_i = (N+1)/2N$, $i = 1,\ldots,N$ and when the whole population is sampled e in (9) unlike e_{HT} in (3) equals $\bar{Y}$!

A generalization of the theorem just stated is obtained by replacing in (II) $y_i - \theta x_i$ by any function

$$\pi_i(y_i, \theta),$$

covering many practical situations including (optimal) estimation of quantiles. The appeal, to the practitioners, of this approach is evident from the fact that special cases of the function ϕ_i above were already in common use (Binder, 1983)

before the present theory (Godambe & Thompson, 1986a) was developed. For further applications we refer to a later paper of Godambe and Thompson (1986b).

References

Basu, D. (1958): On sampling without replacement, *Sankhya* 20, 287-294.

Bellhouse, D. R. (1988): A brief history of random sampling methods, in *Handbook of Statistics*, Volume 6, P. R. Krishnaiah and C. R. Rao (eds.), Neth. Holland, 1-14.

Binder, D. A. (1983): On the variances of asymptotically normal estimators from complex surveys, *Int. Statist. Rev.* 51, 279-292.

Brewer, K. W. R. (1963): Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process, *As. J. Statist.* 5, 93-105.

Brewer, K. W. R. (1979): A class of robust sampling designs for large-scale surveys, *J. Amer. Statist.* 74, 911-915.

Cochran, W. G. (1939): The use of analysis of variance in enumeration by sampling, *J. Amer. Statist. Ass.* 34, 492-510.

Godambe, V. P. (1955): A unified theory of sampling from finite populations, *J. R. Statist. Soc.* 28, 269-278.

Godambe, V. P. (1966): A new approach to sampling from finite populations, *J. R. Statist. Soc.* 3, 310-328.

Godambe, V. P. (1982): Estimation in survey sampling: Robustness and optimality, *J. Amer. Statist. Ass.* 77, 393-406.

Godambe, V. P. (1983): Survey-sampling: Modelling, randomization and robustness—a unified theory view, *Proc. Amer. Statist. Ass.* 26-29.

Godambe, V. P. and Thompson, M. E. (1986a): Parameters of superpopulation and survey population: Their relationships and estimation, *Int. Statist. Rev.* 54, 127-138.

Godambe, V. P. and Thompson, M. E. (1986b): Some optimality results in the presence of nonresponse, *Survey Methodology* 12, 29-36.

Hansen, M. H. and Hurwitz, W. N. (1943): On the theory of sampling from finite populations, *Ann. Math. Statist.* 14, 333-362.

Hartley, H. O. and Rao, J. N. K. (1968). A new estimation theory for sampling surveys, *Biometrika* 55, 547-557.

Horwitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *J. Amer. Statist. Ass.* 47, 663-685.

Neyman, J. (1934). On two different aspects of the representative method: The method of stratified sampling and the method of purposive selections, *J. Roy. Statist. Soc.* 97, 558-625.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models, *Biometrika* 57, 377-387.

Tchuprow, A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations, *Metron* 2, 461-493, 646-680.