ESTIMATION IN SURVEY-SAMPLING: ROBUSTNESS AND OPTIMALITY

V.P. Godambe, University of Waterloo

SUMMARY

The major theoretical development of survey-sampling in recent years seem to be primarily motivated by two key results, namely (1) the non-existence of minimum variance unbiased estimation (Godambe, 1955) and (2) the fact that the likelihood function is independent of the sampling design (Godambe, 1966). While these somewhat negative results received considerable attention (Lindley, 1971; Rao, 1977; Royall, 1976) the positive proposals made along with them remained mostly unnoticed: (I) a criterion of optimality to replace minimum variance was laid down (1955) and (II) a suggestion was made to interpret sampling designs as means to robust inference (1966). In this paper we demonstrate how proposals (I) and (II) can lead to estimation which is both *efficient* under an assumed (prior-probabilistic) model and *robust* against possible departures from the model. The robustness here is in relation to the estimator as well as its efficiency (Box and Tiao, 1962).

The twin concepts of random and purposive sampling were understood intuitively and practiced in sample surveys long before they were formalised. It was also informally understood that randomization, unlike purposive selection, renders inference more or less free of any assumptions, i.e. renders it robust against departures from assumptions. Though possibly purposive sampling is of older origin than the other, in the theoretical development random sampling preceded purposive sampling, as will be clear from the subsequent discussion.

The general thesis of the present paper is this. 'Purposiveness' and 'randomization' are two irreducible components of survey-sampling. They determine the efficiency of ultimate estimation; more precisely, optimality in this respect would be the result of blending an appropriate mode of 'randomization' with corresponding 'purposive' elements. Leaving aside the mathematical definition of optimal estimation for the moment we will here explain the above thesis in its historical context.

Neyman's (1934) forceful formalization was intended to show the superiority of random sampling over purposive selection. He argued that the inference based on a purposive sample becomes baseless in case the (prior) assumptions about the survey population, supporting the purposive selection, go wrong. On the other hand the frequency properties of estimation (such as unbiasedness and minimum variance) obtained through randomization are independent of any such prior assumption. Of course, even Neyman made assumptions of homogeneous strata and the like. But these assumptions tended to be less restrictive than the prior probabilistic assumptions (models) on which purposive sampling was based. For instance, Gini-Galvani's (1929) purposive selection could be appropriate only under the assumptions of a probabilistic model providing the regression of the variate under study on the auxiliary variate.

Now in retrospect it appears that Neyman's thoroughgoing frequency approach using unbiased minimum variance (UMV) estimation based on randomization had intrinsic limitations. Neyman considered estimators which were based on individual labels only through the stratum to which the individuals belonged. But with the subsequent introduction of modes of randomization more sophisticated than stratification such as unequal probability sampling (Hansen and Hurwitz (1944)), estimators based more essentially on the labels (such as Horvitz- Thompson (1952) type estimators) became common. Further, soon after, for the class of all label based estimators the non-existence of UMV estimation was established (Godambe, 1955; Godambe and Joshi, 1965). On the other hand, 'purposive sampling' received new significance and interpretation from subsequent theoretical development. Within the formal survey-sampling model which takes into account individual labels, the likelihood function was found to be independent of the mode of randomization (Godambe, 1966). This implied that once the sample was drawn any inference (estimation) consistent with the likelihood and conditionality principles should be *independent* of whether the sample was drawn at *random* or was drawn *purposively*. Indeed the likelihood (Barnard et al, 1962) and conditionality (Birnbaum 1961) principles have strong appeal even for many non-Bayesian statisticians. The discovery that the likelihood function is independent of the mode of randomization gave rise to formal theories of (prior-probabilistic) *model based inference* (Royall, 1976) and the related purposive selection mentioned earlier. (In contrast to model based inference, the inference which essentially utilizes the mode of randomization is usually referred to in the literature as *design based inference*.)

Hence the theoretical developments thus far, while sharpening and advancing our understanding of the two basic concepts of random and purposive sampling, have also clarified the basic conflict between the two. As indicated earlier the thesis of the present paper is that the conflict between the two basic concepts can be removed by implementing the considerations of 'model based efficiency' and 'robustness under departures from the model' into a single integrated criterion of optimal estimation.

This paper is a logical extension of previous work by Godambe and Thompson (1971, 1976); there the authors considered *general model failure* and showed that in some (very ristrictive) situations model based inference statements could be replaced in the case of model failure, by corresponding (frequency) design based inferences. In this paper the more common phenomenon of *specific departures from the assumed model* is considered. It is shown that appropriate randomizations (sampling designs) provide with large frequency samples for which inferences under the 'model' and under 'departures' are nearly the same.

In this paper we define a model as a class C of prior distributions on the population (size N) individuals, i.e. on R_N. Corresponding to C let $\{e_\xi : \xi \epsilon C\}$ be the class of Bayes estimators for the population total T. Then we define an estimate $\tilde{e}$ *closest* to the class $\{e_\xi : \xi \epsilon C\}$ and show that $\tilde{e}$ is the usual model C-best estimator i.e. for every sample s, in the class of C-unbiased estimators for the population total, $\tilde{e}$ has smallest C-variance. When C is a complete class of distributions, it is shown that no C-unbiased estimator exists unless the sample s is equivalent to the entire population. That is for a sufficiently broad model, C-best estimator does not exist. Hence we define in an analogous manner an estimate $\tilde{e}$ which is *closest* to the class $\{e_\xi : \xi \epsilon C\}$, under a sampling design p: If E_p and ε_ξ denote expectations under the sampling desing p and the prior distribution ξ then it is shown that in the class of estimates e satisfying $E_p \varepsilon_\xi (e-T)=0$, $E_p \varepsilon_\xi (e-T)^2$ is minimized for $e=\tilde{e}_p$ when it exists. Further if C is complete, the class of estimates satisfying $E_p \varepsilon_\xi (e-T)=0$ is identical with the class of design unbiased estimates, i.e. those satisfying $E_p(e-T)=0$. Thus if under the samping design p an estimate $\tilde{e}_p$ closest to the class of Bayes estimates $\{e_\xi : \xi \epsilon C\}$ exists, it satisfies the optimality criterion of Godambe (1955) and Godambe and Joshi (1965). Hence one would like to find a sampling desing p which minimizes $E_p \varepsilon_\xi (p-T)^2$, $\xi \epsilon C$ to get closest possible to the class of Bayes estimates $\{e_\xi : \xi \epsilon C\}$. Now though under completeness of C, $(\tilde{e}_p, p)$ minimizing $E_p \varepsilon_\xi (\tilde{e}_p - T)^2$ does not exist, good approximations satisfying the *near optimality criterion* defined in the paper exist for different models C. Most designs satisfying the criterion of near optimality are stratified(homogeneous in some respects) sampling designs with appropriate inclusion probabilities. The paper discusses many examples illustraing and emphasizing how near optimal sampling designs provide a well defined mechanism for *eliminating* nuisance parameters in contrast to their estimation in the model based approach. The latter estimation can be justified only for very restrictive models; hence the limitations of pure model based approach. Now even statisticians professing to adopt model based approach, recommend drawing sample with suitable randomization for 'balancing' Now this inference based on randomization in model based theory is not only ad hoc but is fundamentally in conflict with the basic principle on which the model based theory is founded--namely the likelihood principle.

REFERENCES

Barnard, G.A., Jenkins, G.M. and Winston, C.B. (1962), "Likelihood Inference and Time Series" (with discussion), *Journal of the Royal Statistical Society,* A, 125, 321-372.

Basu, D. (1971), "An Eassay on Logical Foundations of Survey-Sampling", *Foundations of Statistical Inferences,* Eds. V.P. Godambe and I.A. Sprott, Holt Rinehard and Winston of Canada, 203-242.

Birnbaum, A. (1962), "On Foundations of Statistical Inference", *Journal of the American Statistical Association,* 57, 269-326.

Box, G.E.P. and Tiao, G.C. (1962), "A further look at Robustness via Bayes' Theorem", *Biometrika,* 49, 419-432.

Cassel, C.M., Sarndal, Carl E., Wretman, Jan H., (1976), "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations", *Biometrika,* 63, 615-620.

Gini, C. and Galvani, (1929). Di-una applicazione del metoda rappresentativo all'utimo censimento italiano della popolazione, *Annali di Statistica,* 4, 1-107.

Godambe, V.P. (1955), "A Unified Theory of Sampling from Finite Populations", *Journal of the Royal Statistical Society,* B, 17, 269-278.

Godambe, V.P. (1966), "A New Approach of Sampling from Finite Problems", *Journal of the Royal Statistical Society,* B, 28, 310-328.

Godambe, V.P. and Joshi, V.M. (1965), "Admissibility and Bayes Estimation in Sampling Finite Populations -- I", *Annals of Mathematical Statistics,* 36, 1707-1722.

Godambe, V.P. and Thompson, M.E. L1971), "Bayes, Fiducial and Frequency Aspects of Statistical Inference in Survey-Sampling", *Journal of the Royal Statistical Society,* B, 33, 361-390.

Godambe, V.P. and Thompson, M.E. (1976), "Philosophy of Survey Sampling Practice", *Eds. Harper and Hooker,* D. Reidel Publishing Co., Dordrecht-Holland, 103-123.

Hansen, M.H. and Hurwitz, W.N. (1943), "On the Theory of Sampling from Finite Populations", *Annals of Mathematical Statistics,* 14, 333-362.

Horvitz, D.G. and Thompson, D.J. L1952), "A Generalization of Sampling Without Replacement from a Finite Universe", *Journal of the American Statistical Association,* 47, 662-685.

Lehmann, E.L. (1959). *Testing Statistical Hypotheses.* New York: Wiley.

Lindley, D.V. (1971), "Bayesian Statistics. A Review", *Society for Industrial and Applied Mathematics,* Philadelphia.

Neyman, J. (1934), "On Two Different Aspects of Representative Method" *Journal of the Royal Statistical Society,* 97, 558-606.

Rao, C.R. (1952), "Some theorems on minimum variance unbiased estimation", *Sankhyā,* Ser. A, 12, 27-42.

Rao, C.R. (1965), "*Linear Statistical Inference and Its Applications*", John Wiley and Sons Inc.

Rao, C.R. (1977), "Some Real Problems of Sample Surveys", *Sankhyā,* C, 39, 128-139.

Royall, Richard M. (1976), "Current Advances in Sampling Theory: Implications for Human Observational Studies", *American Journal of Epidemiology,* 104, 463-474.

Royall, R.M. and Herson, J. (1973), "Robust Estimation in Finite Populations I' *Journal of the American Statistical Association,* 68, 880-889.

Scott, A.; Brewer, K.R.W. and Ho, E.W.H. (1978), "Finite Population Sampling and Robust Estimation", *Journal of the American Statistical Association,* 73, 359-361.

Yates, F. (1960), "*Sampling Methods for Censuses and Surveys*", 3rd Edition, London: Griffin.