SURVEY SAMPLING: MODELLING, RANDOMIZATION AND ROBUSTNESS – A UNIFIED THEORY VIEW

V.P. Godambe, University of Waterloo

1. INTRODUCTION. With the usual notation let a finite population of size N with labelled individuals i be denoted by $\mathcal{P} = \{i: i=1,\ldots,N\}$. Let the variate vector under study be $\underset{\sim}{y} = (y_1,\ldots,y_i,\ldots,y_N)$. We want to estimate the population total $T(\underset{\sim}{y}) \doteq \sum_1^N y_i$ or the population mean $T(\underset{\sim}{y})/N$ on the basis of a sample s, $s \subset \mathcal{P}$ and the observed values y_i: $i \in s$. The sample s is drawn from $\mathcal{S} = \{s: s \subset \mathcal{P}\}$ using a probability distribution on $\mathcal{S}$. This probability distribution say p is called a sampling design. We assume that the prior knowledge concerning the population vector $\underset{\sim}{y}$ can be formalized as a class $C(\mathcal{A})$ of distributions ξ on $\mathbb{R}_n$ as follows: For a given vector $\underset{\sim}{a} = (a_1,\ldots,a_N)$

$$C(\underset{\sim}{a}) = \left\{ \xi: \begin{array}{ll} \text{(i)} & y_1,\ldots,y_N \text{ when distributed as } \xi \text{ are probabilistically mutually independent,} \\ \text{(ii)} & \varepsilon_\xi(y_i) = a_i + \beta x_i, \\ \text{(iii)} & \varepsilon_\xi(y_i - a_i - \beta x_i)^2 = \sigma_i^2,\ i=1,\ldots,N \\ & \text{where } \underset{\sim}{x} = (x_1,\ldots,x_N) \text{ is given and } \beta,\ \sigma_i^2,\ i=1,\ldots,N \\ & \text{belong to } \textit{specified} \text{ (known) intervals} \end{array} \right.$$

$\varepsilon_\xi(\cdot)$ denoting the expectation under the distribution ξ. Then

$$C(\mathcal{A}) = \bigcup_{a \in \mathcal{A}} C(\underset{\sim}{a}). \qquad (1.1)$$

We shall elaborate on the intervals of β and σ_i^2 and the set $\mathcal{A}$ later. $\underset{\sim}{x} = (x_1,\ldots,x_N)$ in (1.1) is called the fixed *covariate* vector. $C(\mathcal{A})$ in (1.1) would also be called a *model*. Clearly the model $C(\mathcal{A})$ is constructed from the usual regression model by taking into account, (A) the possible departures from the regression line, by the introduction of the variable vector $\underset{\sim}{a}$ and (B) the possible variations of variances of y_i, $i=1,\ldots,N$. The model $C(\mathcal{A})$ was first studied in detail by Godambe, (1982).

2. MODEL BASED ESTIMATION. If in the sub-model $C(\underset{\sim}{a})$ in (1.1) σ_i^2, $i=1,\ldots,N$ are specified the least square estimate for the population total $T(\underset{\sim}{y})$ is given by

$$e_{\underset{\sim}{a}} = \sum_{i\in s} y_i + \hat{\beta}_{\underset{\sim}{a}} X_{\bar{s}} + \sum_1^N a_i - \sum_s a_i \qquad (2.1)$$

where $X_{\bar{s}} = \sum_{\bar{s}} x_i$, $\sum_s$ and $\sum_{\bar{s}}$ being summations for $i \in s$ and $i \notin s$ respectively; $\hat{\beta}_{\underset{\sim}{a}}$ is the usual least squares estimate for β i.e.

$$\hat{\beta}_a = \left(\sum_s x_i(y_i - a_i)/\sigma_i^2\right) / \left(\sum_s x_i^2/\sigma_i^2\right). \qquad (2.2)$$

It is important to note that the estimate $e_{\underset{\sim}{a}}$ in (2.1) is independent of the sampling design used to draw the sample s. Obviously the estimate $e_{\underset{\sim}{a}}$ is very sensitive to (i) the variations of $\underset{\sim}{a}$ and (ii) the variations in σ_i^2, $i=1,\ldots,N$. For a restrictive situation when $\underset{\sim}{a}$ corresponds to polynomial regression the variation (i) can be dealt with by drawing a suitably balanced sample, (Royall and Herson, 1973). To deal with a very restrictive variation in (ii), Scott, Brewer and Ho (1978) have suggested use of unequal probability sampling designs. Royall and Pfefferman (1982) have suggested use of a random sampling design to draw a sample balanced on unknown factors.

The above suggestions to supplement model based estimation with randomization to achieve robustness are interesting. But they are of very limited practical value and moreover, they contradict the 'likelihood principle' which is *fundamental* for model based estimation (Godambe, 1966, 1982); thus the use of randomization here is *adhoc*.

In the above approach the randomization or sampling design is given a *secondary* role to that of the model: the estimate (2.1) is obtained by considering all linear model-unbiased estimates of T and then choosing from them one that has smallest variance relative to the model. Thus the 'expectations' or 'averages' here are exclusively with respect to the model. On the other hand the Unified Theory (Godambe, 1955, 1982) treats distributions generated by randomization and those given by the model on an *equal* footing to investigate the optimum estimation. This approach, as we shall demonstrate below, provides in relation to the model $C(\mathcal{A})$ in (1.1), a logically far more satisfactory and practically far more widely applicable solution than those provided by the model-based estimation with its adhoc supplement of randomization.

It is clear that randomization by itself provides inferences which are *outside* the scope of parametric model-based theory (Kempthorne, 1977, Godambe, 1982a). In conjunction with parametric models, randomization provides *robust* inference, (Godambe, 1982).

3. UNIFIED THEORY ESTIMATION. The optimum estimation here is defined with respect to the model $C = \{\xi\}$, ξ being a probability distribution on $\mathbb{R}_n$ and a class of sampling designs $D = \{p\}$. For a sampling design p, the inclusion probability for the individual i is defined as $\pi_i(p) = \sum_{s \ni i} p(s)$, $i = 1,\ldots,N$. For brevity we suppress p in $\pi_i(p)$ and denote it by just π_i. For all $p \in D$, we assume $\pi_i > 0$, $i=1,\ldots,N$; otherwise the class of estimates B_p given by (3.2), will be empty. Further we restrict the class D to fixed sample size (=n say) designs. That is, if $\nu(s)$ denotes the size of the sample s then $[\nu(s) \neq n] \to [p(s) = 0,\ p \in D]$. For

a given sampling design $p \in D$ we define a class of estimates $e = e(s, y_i: i \in s)$, a real function with arguments s and $y_i: i \in s$,

$$B_p = \{e: E_p \varepsilon_\xi (e-T) = 0,\ \xi \in C\}. \qquad (3.1)$$

If C is a complete class of distributions (ignoring sets of measure 0 or restricting to discrete distributions) we have

$$B_p = \{e: E_p(e-T) = 0,\ \forall \underset{\sim}{y}\}, \qquad (3.2)$$

(Godambe, 1982). For any *fixed* design $p \in D$, $\tilde{e}_p$ is said to be an *optimum estimator* for the population total T (with respect to the model C) if $\tilde{e}_p \in B_p$ and

$$E_p \varepsilon_\xi (\tilde{e}_p - T)^2 \le E_p \varepsilon_\xi (e-T)^2 \quad \text{for all } e \in B_p,\ \xi \in C. \qquad (3.3)$$

Further a pair $(\tilde{e}, \tilde{p})$ of an estimator $\tilde{e}$ and design $\tilde{p}$ is said to provide *optimum estimation* (w.r.t. C and D) iff

$$E_{\tilde{p}} \varepsilon_\xi (\tilde{e}-T)^2 \le E_p \varepsilon_\xi (e-T)^2,\ e \in B_p,\ p \in D,\ \xi \in C. \qquad (3.4)$$

In (1.1) the class $C(A)$ is complete if $A = \{\underset{\sim}{a}\}$ is an N-dimensional interval, Godambe, 1982). Note $[\varepsilon_\xi(e-T) = 0,\ \xi \in C] \to [e-T = 0,\ \text{a.e.}]$ if C is complete; that is here model based unbiased estimation is non-existent. Surely a small N-interval A in (1.1) does not constitute a big departure from the classical regression model obtained from (1.1) by putting $a_1 = a_2 = \dots = a_N$ (unknown). Yet for such a departure model-unbiased estimation does *not* exist at all! *Nor does model-based estimation contain a built in mechanism which can provide practically satisfactory and well-defined approximations to optimum estimation. Such a mechanism is provided by an appropriate sampling design within the frame-work of the Unified Theory:*

For any sampling design $p \in D$ and any distribution $\xi \in C(A)$

$$E_p \varepsilon_\xi (e-T)^2 \ge \sum_1^N \frac{\sigma_i^2}{\pi_i} - \sum_1^N \sigma_i^2 \qquad (3.5)$$

for any estimate $e \in B_p$, in (3.2), (Godambe and Joshi, 1965). Further for any specific $\underset{\sim}{a}$, β and sampling design $p \in D$, the estimator

$$e^* = \sum_s \frac{y_i - a_i - \beta x_i}{\pi_i} + \sum_1^N (a_i + \beta x_i) \qquad (3.6)$$

is in B_p of (3.2) and for the submodel $C_\beta(\underset{\sim}{a}) \subset C(\underset{\sim}{a})$ obtained by fixing β in (1.1), e^* attains the lower bound in (3.5). That is,

$$\varepsilon_\xi E_p (e^* - T)^2 = \sum_1^N \frac{\sigma_i^2}{\pi_i} - \sum_1^N \sigma_i^2,\ \xi \in C_\beta(\underset{\sim}{a}), \qquad (3.7)$$

(Godambe, 1982). Thus the estimate e^* is locally i.e. for specified $\underset{\sim}{a}$ and β optimum. Now we try to find an estimate, independent of $\underset{\sim}{a}$ and β, which provides a good approximation (in the following well defined sense) to e^* under the model $C(A)$. Let $\bar{e}$ be the estimate obtained from e^* by putting in it $a_i = 0$, $i = 1, \dots, N$ and $\beta = 0$. Then it is easy to see that

$$e^* = \bar{e} - \Delta \qquad (3.8)$$

where

$$\Delta = \sum_s \frac{a_i}{\pi_i} - \sum_1^N a_i + \beta\left(\sum_s \frac{x_i}{\pi_i} - \sum_1^N x_i\right),$$

$$= \varepsilon_\xi \bar{e} - \varepsilon_\xi T,\ \xi \in C_\beta(\underset{\sim}{a}). \qquad (3.9)$$

Further,

$$E_p \varepsilon_\xi (\bar{e}-T)^2 = E_p \varepsilon_\xi (e^*-T)^2 + E_p \Delta^2,\ \xi \in C_\beta(\underset{\sim}{a}). \qquad (3.10)$$

It is also easy to see that

$$E_p \Delta^2 \le \left[\sqrt{E_p\left(\sum_s \frac{a_i}{\pi_i} - \sum_1^N a_i\right)^2} + \beta \sqrt{E_p\left(\sum_s \frac{x_i}{\pi_i} - \sum_1^N x_i\right)^2}\,\right]^2 \qquad (3.11)$$

4. APPLICATIONS. Several important conclusions, with far reaching practical implications follow from (3.5)-(3.11). Let us suppose that the sampling design p employed to select a sample s has inclusion probabilities π_i, $i = 1, \dots, N$ and is such that with respect to the given model $C(A)$ in (1.1)

$$E_p \Delta^2 / \left(\sum_1^N \sigma_i^2/\pi_i - \sum_1^N \sigma_i^2\right), \qquad (4.1)$$

(see (3.8), (3.11)) is sufficiently small. In this situation the use of the estimate $\bar{e}$ is justified, for by (3.8)-(3.11), and (4.1), the expected variance of $\bar{e}$ and that of the locally (i.e. or given $\underset{\sim}{a}$ and β) optimum estimate e^* differ very little; so do the two estimates $\bar{e}$ and e^* themselves differ a little on most probable samples s. Note here the model-based estimate $e_{\underset{\sim}{a}}$ in (2.1) cannot be computed because of the dependence of $\hat{\beta}$ in (2.2) on σ_i^2, $i = 1, \dots, N$ and $\underset{\sim}{a}$ which are not specified. Can the model-based approach suggest some well defined approximation here?

A more realistic situation is obtained when the model (1.1) is a small departure from the usual regression model obtained by putting in (1.1), $a_1 = a_2 = \dots = a_N$ (*unknown*). Here for given inclusion probabilities π_i, $i = 1, \dots, N$ reduce the value of $E_p \Delta^2$ in (4.1), we should choose a sampling design $p \in D$ which provides with large probability samples s for which simultaneously

$$\left|\sum_s x_i/\pi_i - \sum_1^N x_i\right| \text{ and } \left|\sum_s 1/\pi_i - N\right| \text{ are very small.} \qquad (4.2)$$

One way of achieving the balancing in (4.2) is to choose from D a sampling design which, for given π_i, $i = 1, \dots, N$, consists of stratified sampling with strata as homogeneous as possible in x_i and π_i simultaneously. Then use the estimate $\bar{e}$.

For, let the population P be divided into strata, P_j, $i = 1, \dots, k$, $P = \cup_j P_j$ and $s_j = s \cap P_j$ then in (4.2) $\sum_s x_i/\pi_i = \sum_j \sum_{s_j} x_i/\pi_i$.

Further because of the homogeneous stratification in x_i and π_i, $\sum_{s_j} x_i/\pi_i \simeq n_j(s_j)\cdot \sum_{P_j} x_i/\sum_{P_j}\pi_i$ $n_j(s)$ denoting size of s_j. But stratified sampling implies fixed size sample from each stratum that is $n_j(s_j) = n_j = \sum_{P_j}\pi_i$. Hence in (4.2) $\sum_s x_i/\pi_i \simeq \sum_1^N x_i$. Similarly $\sum_s 1/\pi_i \simeq N$.

Finally we discuss the factors that should govern the choice of the inclusion probabilities π_i, $i = 1,\ldots,N$. From (3.7) and (3.11) it follows that if in the model (1.1), β is considerably larger than σ_i^2 and $|a_i|$ we might ignore the factors $\sum_1^N \sigma_i^2/\pi - \sum_1^N \sigma_i^2$ and $E_p(\sum_s a_i/\pi_i - \sum_1^N a_i)^2$ and just minimize $E_p(\sum_s x_i/\pi_i - \sum_1^N x_i)^2$; the minimum value namely 0 is achieved by $\pi_i \propto x_i$ i.e. $\pi_i = nx_i/\sum_1^N x_i$. On the other hand if σ_i^2, $i = 1,\ldots,N$ are much larger than β and $|a_i|$ then (even if their values are only approximately known) we should have $\pi_i \propto \sigma_i$ i.e. $\pi_i = n\sigma_i/\sum_1^N \sigma_i$. Further with $\pi_i \propto \sigma_i$ we should choose a stratified random sampling design with as sharp as stratification in x_i/π_i as is possible, if $|a_i|$ are ignorably small. If σ_i^2 are unknown but are known to be very large, applying some kind of *minimaxity argument** to the factor $\sum_1^N \sigma_i^2/\pi_i - \sum_1^N \sigma_i^2$ in (3.7) and (3.10), we may choose π_i, $i = 1,\ldots,N$ all equal to n/N. Then again use stratified random sampling with as sharp a stratification in x_i/π_i i.e. x_i as is possible. In any case the estimate is $\bar{e}$. If σ_i^2, $i = 1,\ldots,N$ are known then as indicated above we choose a design $p \in D$ with $\pi_i = n\sigma_i/\sum_1^N \sigma_i$ admitting stratified sampling with a sharp stratification in x_i/σ_i, assuming $|a_i|$ are ignorably small. If such a sharp stratification is unavailable but a *plausible* value of β say β_0 is available then we may improve on the estimate $\bar{e}$ by

$$\bar{\bar{e}} = \frac{\sum_1^N \sigma_i}{n} \cdot \sum_s \frac{y_i - \beta_0 x_i}{\sigma_i} + \beta_0 \sum_1^N x_i$$

A minimaxity argument similar to one mentioned before also suggests that in a commonly occurring practical situation when one is required to estimate from the same sample s, population totals of many variates satisfying different regression models with different covariates, one should employ a design with all inclusion probabilities $\pi_i = n/N$, $i = 1,\ldots,N$, involving a sharp stratification on all the covariates.

5. GENERAL REMARKS. One important distinction between the Unified Theory approach and the model-based approach brought out by the above analysis is this: In the former one tries to *eliminate* the nuisance parameters like β, σ_i^2, $\underset{\sim}{a}$ while in the latter one tries to *estimate* them. Evidently the 'elimination' procedure is more widely applicable than the 'estimation' procedure; no model based estimates of σ_i^2 or $\underset{\sim}{a}$ are at all available excepting under a restrictive assumption that all σ_i^2 $(i = 1,\ldots,N)$ are equal and so are all a_i.

It should be clear from the forgoing discussion that the estimation based on the Unified Theory is available for many situations (broader models) for which model based estimation (even with its supplement of randomization) has no solutions. Further the robust estimation provided by the unified theory has *two* components: (i) The sampling design for which in (3.10), $E_p\Delta^2$ is small produces, because of (3.8), more often samples on which the estimates e^* and $\bar{e}$ agree. Thus criterion robustness is achieved. (ii) The small value of $E_p\Delta^2$, by (3.10) implies that $\varepsilon_\xi E_p(e-T)^2$ and $\varepsilon_\xi E_p(e^*-T)^2$ do not differ much. Hence efficiency robustness. A definition of robustness can be satisfactory only if it is satisfied in relation to both criterion and efficiency (Box and Tao, 1962).

Many thanks are due to Mary Thompson for useful discussions. This work was done when the author enjoyed the hospitality of the statistics department of POONA UNIVERSITY (Pune, India) as a fellow. He is grateful to all concerned. The author also would like to thank his hosts GAVASKARS for providing a very comfortable and intellectually stimulating accommodation at Pune (India).

**Theorem*. Let $(\pi_1(P),\ldots,\pi_i(P),\ldots,\pi_N(P))$ be a permutation P of a fixed vector $\pi = (\pi_1,\ldots,\pi_i,\ldots,\pi_N)$. Further let $M(\pi) = \max_P \sum_1^N \sigma_i^2/\pi_i(P)$, σ_i^2 being fixed. Then subject to $\sum_1^N \pi_i = n$ $\mathrm{Min}_\pi M(\pi) = (N/n)\sum_1^N \sigma_i^2$.

REFERENCES

Box, G.E.P. and Tiao, G.C. (1962). A further look at robustness via Bayes' Theorem. Biometrika, 49, 412-32.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. Jour. Royal. Statist. Ser. B. 17, 269-78.

Godambe, V.P. (1966). A new approach to sampling from finite populations. Jour. Royal. Statist. Soc. B. 28, 310-28.

Godambe, V.P. (1982). Estimation in survey sampling: robustness and optimality. J. Ann. Statist. Assoc. 77, 393-403.

Godambe, V.P. (1982). Likelihood principle and randomisation in Statistics and Probability: Essays in Honour of C.R. Rao. Eds. G. Kallianpur, P.R. Krishnaiah, J.K. Ghosh. North-Holland Publishing Company.

Godambe, V.P. (1982)[a]. Ancillarity principle and a statistical paradox. JASA 77, 931-3.

Godambe, V.P. and Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations I. Ann. Math. Statist. 36, 1707-22.

Kempthorne, O. (1977). Why randomise? J. Statist. Plan. and Inf. 1, 1-27.

Royall, R.M. and Herson, J.H. (1973). Robustness estimation in finite populations I. J. Am. Statist. Assoc. 68, 880-9.

Royall, Richard M. and Pfefferman, Dany (1982). Balanced samples and robust Bayesian inference in finite population sampling, Biometrika, 69, 2401-9.

Scott, A., Brewer, K.R.W. and Ho, E.W.H. (1978). Finite population sampling and robust estimation. J. Am. Statist. Assoc. 73, 359-61.