Survey Methodology, June 1990
Vol. 16, No. 1, pp. 3–29
Statistics Canada

History and Development of the Theoretical Foundations of Survey Based Estimation and Analysis

J.N.K. RAO and D.R. BELLHOUSE[1]

ABSTRACT

Early developments in sampling theory and methods largely concentrated on efficient sampling designs and associated estimation techniques for population totals or means. More recently, the theoretical foundations of survey based estimation have also been critically examined, and formal frameworks for inference on totals or means have emerged. During the past 10 years or so, rapid progress has also been made in the development of methods for the analysis of survey data that take account of the complexity of the sampling design. The scope of this paper is restricted to an overview and appraisal of some of these developments.

KEY WORDS: Foundations of inference; Analysis of survey data; Computer software.

1. SOME EARLY MILESTONES

The motivation behind much of the work in survey sampling prior to the 1950's or 60's was the desire to obtain reasonably efficient estimates, at a desired cost, of totals, means, or proportions for large, and increasingly complex-structured, finite populations. A discussion of the early work in sampling human populations may be found in several review papers (see *e.g.*, Hansen, Dalenius and Tepping 1985 and Bellhouse 1988).

The history of the mathematical theory of survey sampling has its origins in the late nineteenth century through the work of the Norwegian statistician A.N. Kiaer. Kiaer was the first to promote what was then called 'the representative method', or sampling, over complete enumeration. What Kiaer (1897) meant by representative sampling was that the sample should mirror the parent finite population. This can be achieved in two ways, by randomization or by balanced sampling through purposive selection. Initially, purposive selection was the preferred method of sample selection, but gradually randomization became a strong competitor to balanced sampling for sample selection. By the 1920's random sampling and purposive selection were both widely used as sample selection techniques. The major theoretical developments in both areas which occurred during this era are summarized in Bowley (1926). This summary includes the development of stratified random sampling with proportional allocation and the derivation of formulae to obtain the precision of an estimate from a purposively selected sample.

The equal footing of random sampling and purposive selection gradually changed after the publication of Neyman's (1934) classic paper. Neyman was able to show, both theoretically and with practical examples, why random sampling was preferable to purposive selection for the large-scale sampling problems of the day. With the publication of the 1934 paper, Neyman also opened up new avenues of development for random sample selection techniques. Previously, Bowley and his followers used only sampling designs with equal inclusion

[1] J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6. D.R. Bellhouse, Department of Statistics, University of Western Ontario, London, Ontario, N6A 5B9.

probabilities for every population unit. Their reasoning was that this method of sampling would provide a representative sample of the universe. Neyman (1934) broke out of this sampling straitjacket with his theories of stratified sampling with "optimal" allocation and cluster sampling with ratio estimation. In both situations, "valid" estimates of population totals, means or proportions are obtained without reliance on a representative sample selected through a design with equal inclusion probabilities. Neyman's final contribution to the theory of survey sampling is his introduction of cost functions to find the sample allocation in two phase sampling which minimized the variance subject to a fixed budget (Neyman 1938).

Neyman's fundamental contributions inspired various important extensions of his theory. Among these, we should mention ratio and regression estimation with two-phase sampling (Cochran 1939), determination of "optimal" stratification points and "optimal" allocation with multiple parameters/characters (Dalenius 1957), and sampling on two occasions with partial replacement of units (Jessen 1942) which was subsequently extended by Patterson (1950) and Hansen *et al.* (1953, pp. 470-503) to sampling on more than two occasions (also called rotation sampling). Rotation sampling and associated "composite" estimates are now extensively used to estimate levels and changes from continuing large scale, multi-purpose surveys (*e.g.*, the Current Population Survey (CPS) carried out by the U.S. Bureau of the Census).

Neyman's work also greatly influenced Morris Hansen, William Hurwitz, and their colleagues at the U.S. Bureau of the Census. Inspired by their practical problems in large-scale survey design and by Neyman's approach to sampling theory, Hansen and Hurwitz (1943) developed the theory of sampling with probability proportional to size and with replacement (also called PPS sampling). The effect of this approach to multistage surveys is that it provides approximately equal interviewer work loads which makes the administration of a multistage survey easier. This procedure also leads to significant reductions in the variances of the estimates, by controlling the variability arising from unequal cluster sizes without actually stratifying by size and thus allowing stratification on other variables to reduce variance. The theory of Hansen and Hurwitz was extended by Horvitz and Thompson (1952) and Narain (1951) to unequal probability sampling without replacement. By making the inclusion probabilities of units at each stage proportional to their sizes, the desirable features of the Hansen-Hurwitz method are retained, using the so-called Horvitz-Thompson estimator of a population total. The basic work of Horvitz and Thompson and Narain stimulated many theoretical and applied contributions to unequal probability sampling without replacement. Brewer and Hanif (1983) and Chaudhuri and Vos (1988) have provided comprehensive accounts of these developments.

Madow and Madow (1944) have given the basic theory of systematic sampling, and introduced population models to examine the features of systematic sampling. Cochran (1946) introduced the "superpopulation" approach in which the finite population is regarded as being drawn from an infinite superpopulation having certain properties. The expected (or anticipated) variances under the superpopulation model are then compared to study the relative efficiency of alternative sampling strategies. His 1946 paper stimulated much subsequent research in the use of superpopulation models in the choice of sampling strategies and also for model-dependent or model-assisted inference (see Section 2).

Mahalanobis (1946) developed the technique of interpenetrating subsamples, and used it extensively in large-scale surveys in India for assessing both sampling and non-sampling errors. This technique consists of drawing the sample in the form of two or more independent subsamples according to the same sampling scheme such that each subsample provides a valid estimate of the parameter of interest. By assigning the subsamples to different interviewers

(or interviewer teams), a valid estimate of the total variance can be obtained that takes proper account of the correlated response variance component due to interviewers. Deming (1960) used this method (sometimes called replicated sampling) extensively to obtain simple estimates of variance. It has led to resampling techniques such as the jackknife, balanced repeated replication and the bootstrap for getting variance estimates of complex non-linear statistics (see Section 3).

Yet another milestone in the emergence of ideas and theory surrounding complex surveys is the concept of design effect (DEFF), due to Leslie Kish (see Kish 1965, sec. 8.2). The design effect is defined as the ratio of the actual variance of a statistic under the specified design to the variance which would be achieved under a simple random sample of the same size. The concept of design effect has been found to be especially useful in the presentation and modelling of sampling errors, and also in the analysis of survey data involving clustering and stratification (see Section 4).

2. THEORETICAL FOUNDATIONS

Although Neyman (1934) and others obtained best linear unbiased estimators for simple designs using the standard Gauss-Markov set-up, the development of traditional sampling theory progressed more or less inductively. Estimators (and designs) which appeared reasonable were considered and their relative properties carefully studied by analytical and/or empirical methods, mainly through comparisons of bias and mean square error, and sometimes also using anticipated mean square error or variance under plausible superpopulation models. As noted by Hansen *et al.* (1983), unbiasedness of estimators under a given design was not insisted on since it "often results in much larger mean square errors than necessary". Instead, asymptotic design consistency of estimators was insisted on, at least when aggregate estimates from reasonably large samples are needed, and the mean square errors of selected asymptotically design consistent estimators were compared to arrive at a suitable estimator (and design). Moreover, in large-scale surveys involving a great many statistics, uniform estimation procedures are often insisted on at the expense of variance inflation for some statistics (compared to alternative estimators tailored to each statistic), due to time, cost and other operational constraints.

Despite the usefulness of the traditional approach, the need for a formal framework for inference from survey data was long felt. Realizing this need, several statisticians have made important contributions to the theoretical foundations of inference from survey data, especially during the past 10-20 years. Several review papers (see *e.g.*, Chaudhuri 1988) and two books (Cassel *et al.*, 1977; Chaudhuri and Vos 1988) discuss various aspects of the theoretical foundations.

Most papers on the theoretical foundations of sampling theory have assumed the following somewhat idealistic set-up. A survey population U consists of N distinct elements identified through the labels $j = 1, \ldots, N$. The characteristic of interest y_j (possibly vector-valued) associated with element j can be known **exactly** by observing element j. Thus response or measurement errors are assumed to be absent or ignored if present. The parameter of interest is the population total $Y = y_1 + \ldots + y_N$ or the population mean $\bar{Y} = Y/N$ (if N is known). A sample is a subset s of U and the associated y-values, *i.e.*, $\{(i,y_i), i \in s\}$, selected according to a sampling plan which assigns a known probability $p(s)$ to s such that $p(s) \geq 0$ for all $s \in S$ (the set of all possible s) and $\sum_{s \in S} p(s) = 1$. The selection probability $p(s)$ can depend on known design variables $z = (z_1, \ldots, z_N)'$, such as stratum indicator variables and size measures of clusters, *i.e.*, $p(s) = p(s \mid z)$ where z_j is possibly vector-valued. For

probability sampling, the inclusion probabilities $\pi_j = \sum_{\{s:j\in s\}} p(s)$ are positive, which permits unbiased or consistent estimation of Y in the traditional sense. It is also customary to impose the condition that the joint inclusion probabilities $\pi_{ij} = \sum_{\{s:(i,j)\in s\}} p(s)$ be positive, which permits unbiased or consistent variance estimation in the traditional sense.

The basic problem is to make inferences (estimation, variance estimation and constructing confidence intervals), about the total Y by observing a sample selected according to a specified sampling plan $p(s)$ and also using available supplementary data. This involves essentially three steps: (i) choice of a sampling plan; (ii) choice of an estimator $\hat{Y}$; (iii) choice of a variance estimator and confidence intervals. There are essentially three different approaches to implement these steps: (i) design-based approach, also called probability sampling approach or randomization approach; (ii) model-dependent approach, also called prediction approach or probability speculation approach (Hájek 1981), (iii) a hybrid approach, called model-based approach or model-assisted approach. Developments to date under each of these three approaches are discussed below.

2.1 Design-based Approach

This approach uses probability sampling both for sample selection and for inference from the data. The probability sampling distribution provides valid inferences irrespective of the population y-values, even in complicated situations, in the sense that the pivotal $t = (\hat{Y} - Y)/s(\hat{Y})$ is approximately $N(0,1)$, at least for large samples, where $s(\hat{Y})$ is the standard error of $\hat{Y}$. This approach has been critized on the grounds that such inferences, although assumption-free, refer to repeated sampling from the survey population involving all samples $s \in S$ and the associated probabilities $p(s)$, instead of just the particular s that has been drawn. This criticism can be countered to some extent by using either conditional design-based inference referring to a subset of S that is "relevant" to the particular s or by a model-assisted approach.

Horvitz and Thompson (1952) made a basic contribution to foundational aspects of design-based inference by formulating three classes of linear estimators of Y, and then raising the possibility that the best (minimum variance) estimator among all possible linear unbiased estimators of Y may not exist, even for simple random sampling. Prompted by the Horvitz-Thompson formulation, Godambe (1955) proposed a general class of linear estimators given by $\hat{Y}_b = \sum_{i\in s} b_{si} y_i$, where the weight b_{si} is attached to element i if s is selected and $i \in s$. He proved that no best unbiased estimator of Y could exist in this class, for any sampling plan $p(s)$. Since the criterion of minimum variance had failed, several alternative criteria for the choice of an estimator were proposed. Among these, the admissibility criterion is of some use but is not sufficiently selective in distinguishing between the merits of estimators since too many estimators are admissible. Ghosh (1987) provides an excellent survey of results on admissibility and related criteria in finite population sampling. New criteria that give rise to a unique choice of estimator in the Godambe class for any sampling plan have also been put forth, but the optimality properties established have questionable relevance (see Rao 1971, Rao and Singh 1973). Basu's (1971) well-known "elephants" example demonstrates the futility of two such criteria, *viz.* necessary bestness and hyperadmissibility.

Godambe (1966) obtained the likelihood function from the sample $\{(i,y_i), i \in s\}$ regarding the N-vector $y = (y_1, \ldots, y_N)'$ as the parameter of interest, but it provides no information on $(y_i: i \notin s)$, and hence on the total Y, since the N population units are essentially treated as N separate post strata. A way out of this difficulty is to ignore some of the data to make the sample non-unique and arrive at an informative likelihood function (Hartley and Rao 1968; Royall 1968). Another route is to combine the uninformative likelihood function with exchangeable priors via Bayes theorem to arrive at informative posterior inferences (Ericson 1969).

Conditional inference has attracted considerable attention (and controversy) in classical statistics since Fisher (1925). The choice of a relevant reference set for making conditional inference is not always clear-cut, but in the context of post-stratification it seems sensible to make design-based inferences conditional on the realized strata sample sizes (Durbin 1969). Holt and Smith (1979) provide the most compelling arguments in favour of conditional design-based inference, although their discussion was confined to post-stratification of a simple random sample. Rao (1985) considered a number of real examples involving random sample sizes to illustrate conditional design-based inference and associated difficulties.

Robinson (1987) considered conditional design-based inference from a simple random sample when only the population total X of a concomitant variable x is known. By conditioning on the observed sample mean $\bar{x}$, he showed that the usual ratio estimator $\hat{Y}_r = (\bar{y}/\bar{x})X$ is conditionally biased. He obtained a conditional bias adjusted ratio estimator given by

$$\hat{Y}_r(adj) = \hat{Y}_r + N(r - b)(\bar{x} - \bar{X})\bar{X}/\bar{x}, \tag{2.1}$$

where $r = \bar{y}/\bar{x}$ and b is the sample regression coefficient. He also showed that a customary variance estimator

$$s_c^2(\hat{Y}_r) = N^2(1 - n/N) \sum_{i \in s} (y_i - rx_i)^2/n(n - 1) \tag{2.2}$$

is conditionally biased, while another classical variance estimator

$$s_a^2(\hat{Y}_r) = (\bar{X}/\bar{x})^2 s_c^2(\hat{Y}_r) \tag{2.3}$$

is in fact conditionally unbiased, for large n. Robinson also showed, through a simulation study, that $s_a^2(\hat{Y}_r)$ is very close to the estimator of conditional variance of $\hat{Y}_r(adj)$.

2.2 Model-dependent Approach

A strict model-dependent approach involves purposive sampling, and the model distribution (generated from hypothetical realizations of $\mathbf{y} = (y_1, \ldots, y_N)'$ obeying the model) provides valid inferences referring to the particular sample s that has been drawn.

The model-dependent approach was first proposed by Brewer (1963) and extensively studied by Royall and his co-workers, starting with Royall (1970). It is best illustrated under a simple regression model

$$E_m(y_i) = \beta x_i, \quad i = 1, \ldots, N; \quad \beta > 0, x_i > 0 \tag{2.4}$$

where E_m denotes the model expectation. It is further assumed that the model variance $V_m(y_i) = \sigma_i^2$ where σ_i^2 is known except for a multiplicative constant, and that the model covariance $\text{cov}_m(y_i, y_j) = 0$, $i \neq j$. Royall (1970) showed that the customary design-unbiased estimator, $N\bar{y}$, under simple random sampling is biased under the model given by (2.4), and that $N\bar{y}$ leads to serious underestimation if the observed sample contains mostly units with small sizes, x_i. These results can also be shown under the conditional design-based approach without assuming a model (Rao 1985).

The best linear model unbiased estimator (or prediction estimator) of Y under the model (2.4) is given by

$$\hat{Y} = \sum_{i \in s} y_i + \sum_{i \in \bar{s}} \hat{\beta} x_i \tag{2.5}$$

which reduces to the usual ratio estimator $\hat{Y}_r$ if $\sigma_i^2 = \sigma^2 x_i$, where $\bar{s} = U - s$ is the set of non-sampled units and $\hat{\beta}$ is the best linear unbiased estimator of β. The uncertainty in $\hat{Y}$ is measured by $E_m(\hat{Y} - Y)^2 = V_m(\hat{Y} - Y)$ which in the case of $\hat{Y}_r$ reduces to

$$V_m(\hat{Y} - Y) = \{X(X - n\bar{x})/(n\bar{x})\}\sigma^2. \tag{2.6}$$

Since (2.6) decreases as $\bar{x}$ increases, the optimal design is a purposive sample consisting of the n units whose x-values are largest, assuming that the population x_i's are known. A model unbiased estimator, $s_m^2(\hat{Y} - Y)$, of $V_m(\hat{Y} - Y)$ is obtained from (2.6) by replacing σ^2 with its weighted least squares estimator $\hat{\sigma}^2$, and the resulting pivotal $t_m = (\hat{Y} - Y)/s_m(\hat{Y} - Y)$ is approximately $N(0,1)$ under the model distribution. These theoretical results are impressive, but such model-dependent strategies could lead to serious biases if the assumed model is not completely correct.

To protect against model misspecifications, Royall and Herson (1973 a,b) considered model deviations consisting of second or higher order polynomial terms in x (say q-th order) or an intercept or both, and demonstrated that a balanced sample for which $\bar{x}^{(j)} = \bar{X}^{(j)}$, $j = 1, \ldots, q$ provides robustness in the sense that $\hat{Y}_r$ remains model unbiased, where $\bar{x}^{(j)} = \sum_{i \in s} x_i^j/n$ and $\bar{X}^{(j)} = \sum_{i \in U} x_i^j/N$. Further, they have shown that stratification on x with optimal allocation and balanced sampling within each stratum together with the separate ratio estimator of Y provides increased efficiency. Purposively chosen balanced samples have a number of difficulties, nevertheless. First, due to lack of rigorous rules in the sample selection one might be tempted to select units whose x_i are close to $\bar{X}$ (in the case of $q = 1$) which can produce an unrepresentative sample if y is positively correlated with x (Yates 1960, p. 40). Second, balancing is sensitive to departures from the polynomial regression model (Madow 1978, p. 320). Balance is required on the alternative model, which may contain higher-order polynomial terms or other variables or both, and the extra variables in the alternative model must be known in advance. Third, balanced sampling is not feasible for surveys with multiple characters of interest since different samples may be required for each variable.

If the extra concomitant variables z in the model are unknown or unmeasured, Royall and Pfeffermann (1982) recommend simple random sampling since it provides "grounds for confidence that the selected sample is not badly unbalanced on z", but more recently Royall and Cumberland (1988) seem to favour some form of restricted randomization: "Many techniques, including restricted randomization, stratification and systematic sampling, can be used to help achieve balanced samples. We are not advocating one scheme over another; . . .". In any case, it appears that most advocates of the model-dependent approach seem to recommend probability sampling in some form, as noted by Smith (1984), and hence the main difference between the probability sampling approach and the model-dependent approach is in the choice of the pivotal involving the estimator $\hat{Y}$ and a measure of its uncertainty.

Despite the above-mentioned limitations, the model-dependent approach is useful for studying the conditional performances of conventional procedures, under different plausible models. For instance, the variance estimator $s_a^2(\hat{Y}_r)$ is consistent with the behaviour of the conditional variance $V_m(\hat{Y}_r - Y)$ under the model (2.4) with $\sigma_i^2 = \sigma^2 x_i$, while $s_c^2(\hat{Y}_r)$ is model-biased (Royall and Eberhardt 1975). The variance estimator $s_a^2(\hat{Y}_r)$ is also robust to deviations from the assumption $\sigma_i^2 = \sigma^2 x_i$.

2.3 Model-assisted Approach

Hansen, Madow and Tepping (1983) illustrated the dangers in using model-dependent strategies even when the model is apparently consistent with the sample data. By introducing

a misspecification to the model (2.4) which is not detectable through tests of significance from samples as large as 400, they showed that the design-based coverage of the confidence intervals derived from the model-dependent pivotal $t_r = (\hat{Y}_r - Y)/s_a(\hat{Y}_r)$ is substantially less than the desired level and that it becomes worse as the sample size increases. The poor performance of t_r was due to the asymptotic inconsistency of the estimator $\hat{Y}_r$ with respect to their stratified random sampling design.

The model-assisted approach considers only asymptotically design consistent estimators $\hat{Y}$ that are also model unbiased under an assumed model. Variance estimators that are consistent for the design variance of $\hat{Y}$ and at the same time model unbiased (at least approximately) for the conditional variance $V_m(\hat{Y} - Y)$ are also constructed. Thus the resulting pivotal leads to valid inferences under an assumed model and at the same time protects against model misspecifications in the sense of providing valid design-based inferences irrespective of the population y-values. However, very little attention has been given to studying conditional design-based properties of model-assisted strategies under model misspecifications.

Godambe (1955) assumed the model (2.4) with $V_m(y_i) = \sigma_i^2$ and $\text{cov}_m(y_i, y_j) = 0, i \neq j$, and obtained a lower bound, $\sum_{i \in U}(1/\pi_i - 1)\sigma_i^2$, to the anticipated variance of any design unbiased linear estimator, $\hat{Y}_b$. He also showed that any fixed sample size plan with $\pi_i = (nx_i)/X$ together with the Horvitz-Thompson estimator, $\hat{Y}_{HT} = \sum_{i \in s} y_i/\pi_i$, attains the lower bound, provided $\sigma_i^2 = \sigma^2 x_i^2$. "Optimal" design unbiased strategies do not exist if $\sigma_i^2 \neq \sigma^2 x_i^2$, and as a result asymptotically optimal strategies were developed by relaxing the restriction to design unbiased estimators and considering asymptotically design-consistent estimators. The generalized regression estimator

$$\hat{Y}_{reg} = \sum_{i \in s} y_i/\pi_i + \hat{\beta}\left(X - \sum_{i \in s} x_i/\pi_i\right) \tag{2.7}$$

for any fixed sample size plan with π_i proportional to σ_i is asymptotically optimal (*i.e.*, the asymptotic anticipated variance attains the lower bound), where $\hat{\beta}$ is a linear model unbiased estimator of β and $E_m E_p(\hat{\beta} - \beta)^2 \to 0$ as $n \to \infty$, where E_p denotes the design expectation (Särndal 1980). In particular, the best model unbiased estimator $\hat{\beta} = (\sum_{i \in s} w_i x_i y_i)/(\sum_{i \in s} w_i x_i^2)$ with $w_i = 1/\sigma_i^2$ may be chosen.

If $\hat{\beta} = (\sum_{i \in s} w_i x_i y_i/\pi_i)/(\sum_{i \in s} w_i x_i^2/\pi_i)$ with $w_i = 1/x_i$ is chosen, then $\hat{Y}_{reg}$ reduces to the simpler form (ratio estimator)

$$\hat{Y}_{reg} = X\hat{\beta} = \sum_{i \in s} g_{si} y_i/\pi_i, \tag{2.8}$$

where $g_{si} = X/(\sum_{i \in s} x_i/\pi_i)$ and g_{si} converges in probability to 1 as $n \to \infty$ (Särndal and Wright 1984). Särndal, Swensson and Wretman (1989) proposed a new variance estimator for estimators $\hat{Y}$ of the form (2.8) which is design consistent and at the same time approximately unbiased for the conditional variance $V_m(\hat{Y} - Y)$. Their variance estimator for $\hat{Y}_{reg}$ is given by

$$s^2(\hat{Y}_{reg}) = \sum_{i<j \in s} (\pi_i \pi_j - \pi_{ij})\pi_{ij}^{-1}(g_{si}\tilde{e}_i - g_{sj}\tilde{e}_j)^2 \tag{2.9}$$

where $\tilde{e}_i = (y_i - \hat{\beta}x_i)/\pi_i$. For simple random sampling, $s^2(\hat{Y}_{reg})$ reduces to $s_a^2(\hat{Y}_r)$, given by (2.3), which was justified under the prediction and conditional randomization approaches. Kott (1987) proposed a ratio adjustment to the conventional Yates-Grundy variance estimator, $s_{YG}^2(\hat{Y})$, of any model unbiased asymptotically design consistent estimator $\hat{Y}$. His variance estimator

$$\tilde{s}^2_{YG}(\hat{Y}) = s^2_{YG}(\hat{Y})\,[V_m(\hat{Y} - Y)/E_m s^2_{YG}(\hat{Y})] \tag{2.10}$$

is model unbiased and at the same time asymptotically design consistent. However, for estimators of the form (2.8) Särndal *et al.* variance estimator appears simpler since it is obtained simply from the conventional variance estimator $s^2_{YG}(\hat{Y})$ by changing $\tilde{e}_i$ to $g_{si}\tilde{e}_i$.

The conventional regression estimator is obtained by first considering a fixed constant B in place of $\hat{\beta}$ in (2.7), and then substituting a consistent estimator of B_{opt}, the value of B minimizing the design variance. This estimator does not depend on the validity of any model. However, the optimal design variance can be approximately attained in the model- assisted framework by modifying the model (2.4) to $E(y_i) = \beta x_i + \gamma\pi_i$ and then using $(\tilde{\beta},\tilde{\gamma})'$, the weighted regression estimator of $(\beta,\gamma)'$ with weights $w_i = 1/\pi_i^2$. The resulting estimator of Y reduces to (2.7) with $\hat{\beta}$ changed to $\tilde{\beta}$ (Isaki and Fuller 1982; Montanari 1987). Any other choice of $\hat{\beta}$ in (2.7) will give a larger asymptotic design variance.

Little (1983) argued that only models that yield asymptotically design consistent, best linear model unbiased estimators should be used since the latter estimators are optimal if the model is in fact true. One way to accomplish this is by introducing an additional auxiliary variable $u_i = \sigma_i^2(1 - \pi_i)/\pi_i$ into the model (2.4), *i.e.* by using $E(y_i) = \beta x_i + \gamma u_i$ (Särndal and Wright 1984). If we change the model to $E(y_i) = \beta x_i + \gamma\sigma_i^2/\pi_i + \delta\sigma_i^2$ by adding two auxiliary variables σ_i^2/π_i and σ_i^2 to the model (2.4), then we get an asymptotically design consistent, best linear model unbiased estimator of the form $\hat{Y} = \sum_{i\in s} g_{si}y_i/\pi_i$ (Särndal and Wright 1984). The lower bound to asymptotic anticipated variance is also attained if we choose a sampling plan with π_i proportional to σ_i. The above desirable properties, however, are obtained at the expense of a slight increase in the model variance under the original model (2.4).

Godambe and Thompson (1986) employed the theory of estimating functions to derive design consistent estimators through an assumed model. For example, if y_i is expected to be unrelated to π_i for some character y in a multisubject survey, then the "optimal" estimating function gives the Hájek (1971) estimator of $\bar{Y}$:

$$\hat{\bar{Y}}_H = \left(\sum_{i\in s} y_i/\pi_i\right)\Big/\left(\sum_{i\in s} 1/\pi_i\right). \tag{2.11}$$

The superpopulation model here is given by $y_i = \theta + \epsilon_i$, with independent errors ϵ_i, which reflects the situation at hand. The estimator $\hat{\bar{Y}}_H$ avoids the difficulties associated with the Horvitz-Thompson estimator $\hat{Y}_{HT}/N$, as illustrated by the "elephants" example of Basu (1971). The method of estimating functions looks promising, but further work remains to be done on its use in getting "better" estimators or pivotals or both. It is interesting to note that the well-known Fieller method of computing confidence limits for a ratio (Fieller 1932) and the method of Woodruff (1952) for computing confidence limits for medians are essentially equivalent to the method of estimating functions.

The results in Sections 2.2 and 2.3 use models appropriate to unistage sampling. In the case of multistage sampling, the models are more complex due to intra-cluster correlations (Scott and Smith 1969; Montanari 1987). The resulting best linear model unbiased estimators or prediction estimators involve weighted combinations of estimators, where the weights depend on intra-cluster correlations which can be estimated from the sample data. Bellhouse and Rao (1986) investigated the relative efficiency of such estimators, under the repeated sampling framework. Their empirical results suggest that the prediction estimators may not be significantly more efficient than the customary estimator in two-stage sampling with PPS sampling of clusters and simple random sampling within sampled clusters.

If the clusters are regarded as strata and if the strata means are the parameters of interest as in small area estimation, then the prediction estimators of strata means are likely to be significantly more efficient than the customary design-based estimators since the prediction estimators "borrow strength" from all the strata unlike the customary estimators. In the case of two-stage sampling with cluster means as parameters of interest, only a prediction estimator for the nonsampled clusters can be implemented.

3. VARIANCE ESTIMATION AND CONFIDENCE INTERVALS

3.1 Linear Statistics

A substantial part of traditional sampling theory is devoted to the derivation of mean square errors or variances of linear estimators of a total Y, and their estimators. Rao (1979) developed a unified approach for estimators belonging to Godambe's general linear class, $\hat{Y}_b = \sum_{i\epsilon s} b_{is} y_i$, which enables the derivation of mean square error in a straightforward fashion, and also exhibits the necessary form of any non-negative quadratic unbiased estimator of the mean square error. For multistage designs, a general estimator of Y is of the form $\hat{Y}_{bm} = \sum_{i\epsilon s} b_{is} \hat{Y}_i$, where s now denotes a sample of primary sampling units (psu's) and $\hat{Y}_i$ is an unbiased linear estimator of psu total Y_i based on subsampling the psu. Unified variance formulae for multistage designs have been worked out by Raj (1966) and Rao (1975).

Large scale surveys often employ many strata, L, with relatively few psu's n_h, sampled within each stratum h. In fact, it is a common practice to select $n_h = 2$ psu's within each stratum to permit maximum degree of stratification of psu's consistent with the provision of a valid variance estimator. If the psu's are sampled with replacement with probabilities p_{hi} in stratum h, then the estimator of total Y is given by $\hat{Y} = \sum_h \bar{r}_h$, and an unbiased variance estimator is simply obtained as

$$s^2(\hat{Y}) = \sum_h \left\{ \sum_i (r_{hi} - \bar{r}_h)^2 \Big/ [n_h(n_h - 1)] \right\}, \tag{3.1}$$

where $\bar{r}_h = \sum_i r_{hi}/n_h, r_{hi} = \hat{Y}_{hi}/p_{hi}$ and $\hat{Y}_{hi}$ is an unbiased estimator of the i-th psu total in stratum $h (i = 1, \ldots, n_h; h = 1, \ldots, L)$. This stratified design is frequently used in comparing methods for nonlinear statistics (Section 3.2). Because of its simplicity, $s^2(\hat{Y})$ is often used even when the psu's are sampled without replacement. This procedure leads to overestimation of variance, but the relative bias would be small if the first stage sampling fraction is small.

3.2 Non-linear Statistics

Many non-linear, finite population parameters of interest, θ, such as ratio, regression and correlation coefficients, can be expressed as smooth functions, $g(\mathbf{Y})$ of totals $\mathbf{Y} = (Y_1, \ldots, Y_q)'$ of suitably defined variates such that $g(\mathbf{Y}) \propto g_1(Y_1/M, \ldots, Y_{q-1}/M)$, where $Y_q = M$, the population size. The parameter θ is estimated by $g(\hat{\mathbf{Y}}) \propto g_1(\hat{Y}_1/\hat{M}, \ldots, \hat{Y}_{q-1}/\hat{M})$. Such estimators are well-behaved even when the variates attached to the elements t are not related to the inclusion probabilities $\pi_t (t = 1, \ldots, M)$ since $g(\hat{\mathbf{Y}})$ is a function only of the Hájek-type estimators $\hat{\bar{Y}}_j = \hat{Y}_j/\hat{M}$ of the means $\bar{Y}_j$. As an example of $g(\hat{\mathbf{Y}})$, the estimator of a finite population regression coefficient $B = \sum (x_t - \bar{X})(y_t - \bar{Y}) / \sum (x_t - \bar{X})^2$ can be written as

$$\hat{B} = [\hat{Z}/\hat{M} - (\hat{X}/\hat{M})(\hat{Y}/\hat{M})] [\hat{W}/\hat{M} - (\hat{X}/\hat{M})^2]^{-1}, \tag{3.2}$$

where $\hat{X}$, $\hat{Z}$ and $\hat{W}$ are the estimators of the totals X, Z and W of the variates x_t, $z_t = y_t x_t$ and $w_t = x_t^2$ respectively.

Variance estimation methods for non-linear statistics, $g(\hat{\mathbf{Y}})$, include the well-known linearization method and resampling techniques like the jackknife, balanced repeated replication (BRR) and the bootstrap. The linearization method is applicable to general sampling designs, but it involves a separate variance formula for each statistic. On the other hand, resampling methods use a single variance formula for all statistics. The jackknife and BRR, however, are strictly applicable only to those designs in which the psu's are sampled **with** replacement (or the first-stage sampling fractions are negligible). The bootstrap seems to be more generally applicable, but it is computationally more cumbersome and its properties have not yet been fully examined.

Linearization method

If we denote the variance estimator of $\hat{Y} = \hat{Y}(y_t)$ for a general design as $v(y_t)$, the linearization method provides a variance estimator for a nonlinear statistic $\hat{\theta}$ as $v(z_t)$ for a suitably defined synthetic variable z_t which depends on the form of $\hat{\theta}$. For a general statistic $\hat{\theta} = g(\hat{\mathbf{Y}})$, the variance estimator is given by

$$s_L^2(\hat{\theta}) = v(z_t) \quad \text{with} \quad z_t = \sum_i y_{ti} g_i(\hat{\mathbf{Y}}), \tag{3.3}$$

(Woodruff 1971), where y_{ti} is the value of ith character for tth unit, and $g_i(\hat{\mathbf{Y}})$ is the partial derivative $\partial g(\mathbf{Y})/\partial Y_i$ evaluated at $\mathbf{Y} = \hat{\mathbf{Y}} (i = 1, \ldots, q)$. One drawback of the formula (3.3) is that the evaluation of partial derivatives may be difficult in some cases, although useful approximations to the desired partial derivatives can be obtained using numerical methods (Woodruff and Causey 1976). The variance estimator can also be obtained in many cases, without actually evaluating the partial derivatives g_i, by recasting $\hat{\theta}$ as a ratio-type statistic and using the usual variance formula for a ratio. For example, the sample regression coefficient $\hat{B}$ may be expressed as $\hat{B} = \hat{Y}(z_{1t})/\hat{Y}(z_{2t})$ with $z_{1t} = (y_t - \hat{\bar{Y}})(x_t - \hat{\bar{X}})$ and $z_{2t} = (x_t - \hat{\bar{X}})^2$, so that

$$s_L^2(\hat{B}) = v(z_{1t} - \hat{B} z_{2t})/[\hat{Y}(z_{2t})]^2. \tag{3.4}$$

Similar techniques can be used for other statistics like the multiple regression coefficients (Fuller 1975; Folsom 1974). Binder (1983) extended the scope of linearization method to statistics defined implicitly as the solution of a set of nonlinear equations. His formulation covers finite population parameters derived from generalized linear models which include the linear regression model and the logistic regression model.

Resampling methods

We now turn to resampling methods for the commonly used stratified multistage design of Section 3.1. Letting $\hat{\theta}^{hi}$ be the estimator of θ computed from the sample $\{\mathbf{r}_{hi}\}$ after omitting $\mathbf{r}_{hi} = \hat{\mathbf{Y}}_{hi}/p_{hi}$, a jackknife variance estimator of $\hat{\theta} = g(\sum \bar{\mathbf{r}}_h)$ is given by

$$s_J^2(\hat{\theta}) = \sum_h \{(n_h - 1)/n_h\} \sum_i (\hat{\theta}^{hi} - \hat{\theta})^2. \tag{3.5}$$

Several variations of (3.5) can be obtained; for instance, $\hat{\theta}$ in (3.5) may be replaced by $\hat{\theta}^h = \sum_i \hat{\theta}^{hi}/n_h$.

McCarthy (1969) proposed the BRR method for the important special case of $n_h = 2$. A set of J "balanced" half-samples is formed by deleting one psu in the sample from each stratum. This set may be constructed from Hadamard matrices. The BRR variance estimator is given by

$$s^2_{\text{BRR}}(\hat{\theta}) = \sum_j (\hat{\theta}^{(j)} - \hat{\theta})^2/J, \tag{3.6}$$

where $\hat{\theta}^{(j)}$ is the estimator computed from the j-th half sample. Again, several variations of (3.6) can be obtained. The BRR method has been extended recently to the general case of unequal n_h, using asymmetrical orthogonal arrays (Gupta and Nigam 1987; Wang and Wu 1988).

The bootstrap method for the stratified design involved the following steps (Rao and Wu 1988): (i) Draw a simple random sample $\{\mathbf{r}^*_{hi}\}_{i=1}^{m_h}$ of size m_h with replacement from $\{\mathbf{r}_{hi}\}_{i=1}^{n_h}$, independently for each h. Calculate

$$\bar{\mathbf{r}}_{hi} = \bar{\mathbf{r}}_h + [m_h/(n_h - 1)]^{1/2}(\mathbf{r}^*_{hi} - \bar{\mathbf{r}}_h), \bar{\mathbf{r}}_h = n_h^{-1} \sum_i \mathbf{r}_{hi}$$

and $\bar{\theta} = g(\sum \bar{\mathbf{r}}_h)$. (ii) Independently replicate step (i) a large number, B, of times and calculate the corresponding estimators $\bar{\theta}^1, \ldots, \bar{\theta}^B$. (iii) The bootstrap variance estimator of $\hat{\theta}$ is given by

$$s^2_{\text{BOOT}}(\hat{\theta}) = \sum_b (\bar{\theta}^b - \hat{\theta})^2/(B - 1). \tag{3.7}$$

Confidence intervals can also be obtained by approximating the distribution of $t = (\hat{\theta} - \theta)/s_J(\hat{\theta})$ by its bootstrap counterpart $\bar{t} = (\bar{\theta} - \hat{\theta})/s^*_J(\bar{\theta})$, where $s^*_J(\bar{\theta})$ is obtained from $s^2_J(\bar{\theta})$ by jackknifing the particular bootstrap sample $\{\mathbf{r}^*_{hi}\}$. Two-sided $1 - \alpha$ level "bootstrap-t" confidence intervals on θ are then given by

$$[\hat{\theta} - \bar{t}_{\text{UP}} s_J(\hat{\theta}), \hat{\theta} - \bar{t}_{\text{LOW}} s_J(\hat{\theta})], \tag{3.8}$$

where $\bar{t}_{\text{LOW}}$ and $\bar{t}_{\text{UP}}$ are the lower and upper $\alpha/2$ points of $\bar{t}$ obtained from the bootstrap histogram of $\bar{t}^1, \ldots, \bar{t}^B$. One-sided confidence intervals can also be obtained from the bootstrap histogram. Also, one could use the linearization variance estimator instead of the jackknife variance estimator in constructing the confidence intervals. For confidence intervals we need a much larger number, B, of bootstrap samples than for variance estimation. Regarding the choice of bootstrap sample sizes m_h, the choice $m_h = n_h - 1$ is attractive since it gives $\bar{\mathbf{r}}_{hi} = \mathbf{r}^*_{hi}$.

Comparison of the methods

Theoretical properties of the methods reported in the literature include the following: (1) All the variance estimators reduce to the "standard" one, $s^2(\hat{Y})$ given by (3.1), in the linear case $g(\mathbf{Y}) = Y$. (2) For smooth functions $g(\mathbf{Y})$, all the variance estimators are asymptotically design consistent (Krewski and Rao 1981). The jackknife variance estimator, however, is known to be inconsistent for nonsmooth functions like the quantiles, even in the case of simple random sampling. Hence, caution should be exercised in using jackknife software. (3) If $n_h = 2$ for all h, then the jackknife and linearization variance estimators are asymptotically equal to high order terms for smooth functions $g(\mathbf{Y})$, indicating that the choice between

these methods in this important special case should depend more on other considerations like computational costs (Rao and Wu 1985). Turning to empirical studies, Kish and Frankel (1974) studied the linearization, jackknife and BRR methods, using data from the Current Population Survey and sample designs with $n_h = 2$ clusters from each of $L = 6$, 12 and 30 strata. They evaluated the empirical coverage probability of the $1 - \alpha$ level confidence intervals, $\hat{\theta} \pm t_{\alpha/2} s(\hat{\theta})$, for ratios, regression and correlation coefficients, where $t_{\alpha/2}$ is the upper $\alpha/2$-point of a t-variable with L degrees of freedom and $s^2(\hat{\theta})$ is anyone of the variance estimators. The BRR method performed consistently better, in terms of coverage probability, than the jackknife which in turn was better than the linearization method; the observed differences were small for ratios. The methods performed in the reverse order with regard to stability of variance estimator. Other empirical studies in the literature reported similar results. Regarding the bootstrap, a simulation study by Kovar, Rao and Wu (1988) indicates that the bootstrap t-intervals track the nominal error rate in each tail better than the intervals based on the normal approximation to $t = (\hat{\theta} - \theta)/s(\hat{\theta})$, but the bootstrap variance estimators are less stable than those based on the linearization or the jackknife. The second order equivalence of the latter two variance estimators for the special case $n_h = 2$ is also confirmed.

Computationally simpler methods of variance estimation than the previous methods have also been proposed in the literature, *e.g.*, random group method and partially balanced repeated replication, but these variance estimators do not reduce to the "standard" one in the linear case. Methods of constructing models from which sampling errors can be imputed have also been proposed. Such methods are useful in producing "smoothed" standard errors for estimators for which direct computations have not been made, and also in presenting standard errors in a concise form (*e.g.*, graphs) in published reports.

Wolter's (1985) book gives an excellent introduction to recent developments in variance estimation, and illustrates the methods on data from a variety of large-scale surveys. Recent review papers on variance estimation include Rust (1985) and Rao (1988).

4. ANALYSIS OF SURVEY DATA

Standard methods of data analysis are, in general, based on the assumption of simple random sampling. These methods have also been implemented in standard statistical packages, including SPSS^X, BMDP and SAS. Application of standard methods to survey data without some adjustment for survey design, however, can lead to erroneous inferences, since most such data are obtained from complex sample surveys involving clustering, stratification and unequal probability sampling, and as a result do not satisfy the assumption of simple random sampling. In particular, standard errors of parameter estimates and associated confidence intervals can be seriously understated if the effect of design is ignored in the analysis of data. Similarly, the actual type I error rates of tests of hypotheses can be much bigger than the nominal levels. Standard exploratory data analyses, such as residual analysis to detect model deviations, are also affected. Kish and Frankel (1974) and others drew attention to some of these problems with standard methods and emphasized the need for new methods that take proper account of the complexity of survey data. During the past 10 years or so, rapid progress has been made in developing such methods for the following types of analyses: (a) analysis of multi-way contingency tables; (b) analysis of domain means or domain proportions; (c) linear regression analysis; (d) multivariate analysis including principal component analysis and factor analysis. A brief account of some of these developments is given in this section, and the reader is referred to review articles by Nathan (1988), Rao (1987) and Smith (1984), and a book edited by C.J. Skinner, D. Holt and T.M.F. Smith (1989).

4.1 Analysis of Multi-way Contingency Tables

Chi-squared tests (or likelihood ratio tests) are frequently used for the evaluation and selection of parsimonious models on $\mathbf{p}$, the population cell probabilities, in a multi-way contingency table with T cells. For this purpose, loglinear models are convenient because of their close similarity to analysis of variance in systematically providing test statistics of various hypotheses associated with a multi-way table. Rao and Scott (1984) made a systematic study of the impact of survey design on the standard chi-squared test of goodness-of-fit of a loglinear model, denoted by X^2. They showed that X^2 is asymptotically distributed as a weighted sum, $\sum \delta_i W_i$, of $T - r - 1$ independent χ^2_1 variables W_i, where the weights δ_i are the eigenvalues of a "generalized design effects" matrix and $T - r - 1$ is the degrees of freedom. This general result shows that the survey design can have a substantial impact on the type I error rate of X^2. For instance, under a constant design effects clustering model, $\delta_i = \lambda$ for all i, the actual type I error rate, for nominal level α, is approximately given by $Pr[\chi^2_{T-r-1} > \lambda^{-1}\chi^2_{T-r-1}(\alpha)]$ which increases with the clustering effect, λ.

Rao and Scott (1984,7) obtained simple first-order corrections to X^2 which can be computed from published tables that include estimates of design effects (or standard errors) for cell estimates $\hat{\mathbf{p}}$ and their marginal totals, thus facilitating secondary analyses (see also Fellegi 1980, Gross 1984, and Bedrick 1983). A first-order correction refers $X^2/\hat{\delta}.$ to χ^2_{T-r-1}, where $\hat{\delta}.$ is an estimate of the average design effect $\delta. = \sum \delta_i/(T - r - 1)$ or an estimate of an upper bound on $\delta.$. The corrected test is asymptotically valid in the case of constant design effects clustering, and in general it should perform well when the variability of the δ_i's is small. More accurate, second-order corrections that take account of the variability in the δ_i's can also be obtained by using the Satterthwaite approximation to the weighted sum of independent χ^2 variables (Rao and Scott 1984). These tests, however, require the knowledge of a full estimated covariance matrix of $\hat{\mathbf{p}}$. Alternative methods that take account of the survey design include the Wald statistics based on weighted least squares (Koch, Freeman and Freeman 1975) and the jackknife chi-squared tests (Fay 1985). The latter tests are applicable to survey designs permitting the use of a replication method, such as the jackknife or the BRR. The Wald tests require the full estimated covariance matrix of $\hat{\mathbf{p}}$, whereas the jackknife tests require access to cluster-level estimates.

Fay (1985) and Thomas and Rao (1987) showed that the Wald test which refers to χ^2_{T-r-1}, although asymptotically correct, can become highly unstable as the number of cells in the multi-way table increases and the number of sample clusters decreases, leading to unacceptably high type I error rates compared to the nominal level, α. On the other hand, Fay's jackknife tests and the Rao-Scott corrections performed well under quite general conditions. A simple modification to the Wald test which refers to an F distribution on $T - r - 1$ and $f - T + r + 2$ degrees of freedom performed better than the Wald test in controlling the type I error rate, where f is the degrees of freedom for estimating the covariance matrix of $\hat{\mathbf{p}}$.

4.2 Analysis of Domain Means or Domain Proportions

Analysis of domain (or subpopulation) proportions associated with a binary response variable is of considerable interest to researchers in social and health sciences, and other subject matter areas. Logistic regression models are extensively used for this purpose in conjunction with standard statistical methods for binomial proportions. Rao and Scott (1987) obtained simple first-order corrections to standard chi-squared tests of goodness-of-fit and of nested hypotheses which can be computed from published tables that include estimates of design effects (or standard errors) of domain proportions. Roberts, Rao and Kumar (1987) derived more

accurate second-order corrections to standard tests, but these require access to a full estimated covariance matrix of domain proportions. Diagnostics for detecting outlying domain proportions and influential points in the factor space were developed as well, again taking the sampling design into account.

Koch, Freeman and Freeman (1975) used weighted least squares methods to analyze domain means of a quantitative variable, y, and developed Wald tests of goodness-of-fit of the model and of linear hyptotheses on the model parameters. The performance of Wald tests can be improved, as in Section 4.1, by using an F-modification.

4.3 Linear Regression Analysis

In Section 3.2, we considered design-based inferences on nonlinear, finite population parameters such as the finite population simple regression coefficient B. The pivotal $t = (\hat{B} - B)/s(\hat{B})$ is approximately $N(0,1)$, where $\hat{B}$ is the design-consistent estimator, (3.2), of B, and its standard error, $s(\hat{B})$, can be obtained either through the linearization method as in (3.4) or by using one of the replication methods. This approach readily extends to multiple regression coefficients. The design-weighted estimator $\hat{B}$ or its multiple regression analogue can be obtained by the weighted regression option of standard packages by using the survey weights attached to the sample elements as the weights in the regression. However, the standard error of $\hat{B}$ resulting from this routine remains incorrect.

Some people argue that most users are concerned with inferences on parameters of an appropriate superpopulation model rather than inferences on finite population parameters like B. However, the interest in B can also be justified by considering it as the least squares estimator of the superpopulation parameter β in the model

$$y_i = \alpha + \beta x_i + \epsilon_i \text{ with } E_m(\epsilon_i) = 0, \quad i = 1, \ldots, N. \tag{4.1}$$

If the population size is large, then estimating B is effectively equivalent to estimating β, while if the model (4.1) is misspecified to the extent of making β meaningless, then B may still be of interest as the slope of the least squares line fitted to the N-pairs (y_i, x_i) (Godambe and Thompson 1986).

Scott and Holt (1982) used a model-dependent approach to investigate the effect of two-stage sampling on standard regression analysis. They assumed a regression model of the form (4.1) with equi-correlated error terms ϵ_i within each cluster, as in Fuller (1975). This model also holds for the sample pairs (y_i, x_i), $i \in s$, if the selection probabilities are not related to the dependent variable, as in the case of two-stage random sampling. The results of Scott and Holt indicate that the effect of a positive intra-cluster correlation is to understate the standard errors of parameter estimates, and consequently inflate the type I error rates of customary tests. Wu, Holt and Holmes (1988) made a systematic study of the effect of two-stage sampling on the customary F-statistic, and proposed a correction for the F test for unknown intra-cluster correlation, as an alternative to iterative generalized least squares (GLS) procedure. Both the GLS procedure and the F-correction require known cluster labels which may not be available when the survey data are used for secondary analysis.

If the regression model includes all the design variables z related to the dependent variable, such as stratum indicator variables and size measures of units, and the errors ϵ_i are independent with a constant variance σ^2, then standard regression analysis is valid under the model-dependent approach (Pfefferman and Smith 1985). However, such models may involve too many parameters to be useful. Also, the design variables may not be of intrinsic interest to the user, or may not be available in secondary analysis. In such situations, we are often interested

in models of the form (4.1), where x is not a design variable. The sample pairs (y_i, x_i), $i \in s$ however, may not satisfy the model due to sample selection bias. Nathan and Holt (1980) proposed an adjusted regression approach to take account of selection bias, and compared it with ordinary least squares and the design based approach based on $\hat{B}$ and $s(\hat{B})$. This approach assumes specific relationships between the regression variables and the design variables. Their empirical results indicate that ordinary least squares inferences can be highly unreliable, that the design-based approach is basically reliable except under extreme selection schemes, and that the adjusted regression approach performs well. Pfefferman and Holmes (1985) study the robustness of these procedures to misspecification of relationships between the regression variables, and conclude that the adjusted regression approach is very sensitive to model misspecification. The design-weighted estimator $\hat{B}$ is robust, but a more efficient estimator is obtained by modifying the adjusted regression estimator to be design-consistent for the finite population regression coefficient, B.

4.4 Multivariate Analysis

The methods in Section 4.2 for the analysis of domain means can be extended to the multivariate case of domain mean vectors, but no detailed studies of such extensions have been reported in the literature. The literature on multivariate anlaysis of survey data is largely devoted to the analysis of covariance structures, in particular to principal component analysis and factor analysis. Bebbington and Smith (1977), Tortora (1980) and Skinner, Holmes and Smith (1986) investigated the effect of sample design on standard principal component analysis. Their results indicate that the application of standard methods, without some adjustment for the sample design, can lead to erroneous inferences. In particular, the estimators of eigenvalues and eigenvectors of the covariance matrix, Σ_y, can be severely biased for non-self-weighting sample designs. Skinner, Holmes and Smith (1986) proposed maximum likelihood (ML) estimators, under a multivariate normal model, and probability-weighted (or design-based) estimators, to adjust for the effects of the sample design. Their simulation study indicates that both estimators perform well unconditionally, while the probability-weighted estimators exhibit a conditional model bias. The ML estimators, however, may be sensitive to model misspecification. A probability-weighted version of the ML estimators may be more robust, as demonstrated by Pfefferman and Holmes (1985) in the context of the adjusted regression approach (section 4.3). Fuller (1987) derived design-based estimators of the parameters in factor analysis, and the estimated covariance matrix of the estimators. He showed that the estimated variances based on normal theory can seriously underestimate the true variances of the factor estimators.

5. COMPUTER SOFTWARE

Several computer package programs for variance estimation in complex surveys were developed in the mid to late 1970's, often in conjunction with programs for regression analysis of survey data. Wolter (1985, pp. 393-412) reviewed the latest versions of these programs to about 1985. Among the programs listed by Wolter, the ones most commonly used are CLUSTERS (Verma and Pearce 1977), the programs &PSALMS and &REPERR in the OSIRIS IV system (Vinter 1980 and Lepkowski 1982), SUDAAN (Shah 1981a, 1981b, 1982 and Holt 1979), HESBRR (Jones 1983) and SUPER CARP (Hidiroglou, Fuller and Hickman 1980). The programs HESBRR and the OSIRIS IV program &REPERR use balanced repeated replication as the variance estimation technique; the remaining three use the Taylor linearization method.

Cohen, Burt and Jones (1986) evaluated the variance estimation programs for means and ratios, with the exception of CLUSTERS, using a large data set from the National Medical Care Expenditure Survey. They found that the programs SESUDAAN and RATIOEST in the SUDAAN collection were the most efficient in terms of CPU time usage and easier to program than the others.

One major current trend in software development is the development of menu-driven packages on micro-computers. Variance estimation and specialized survey analysis software is no exception to this trend. A notable enhancement to the commonly used variance estimation programs since 1985 is the introduction of PC CARP (Schnell *et al.* 1986 and Schnell *et al.* 1988), available on IBM AT/XT or compatible micro-computers with a math co-processor. This package, like its predecessor SUPER CARP, uses Taylor linearization methods for variance estimation. A second variance estimation package is also available on micro-computers. The package listed as BELLHOUSE in Wolter (1985, p. 399) has been adapted for IBM micros with or without a co-processor by Rylett and Bellhouse (1988) under the program name TREES. This software uses tree structures to mimic the structure of stratified multistage sampling designs and applies tree traversal algorithms, in conjunction with general results on variance estimation in multi-stage sampling (see section 3.1), to the calculation of variance estimates.

A second trend in the computer implementation of survey variance estimation and survey analysis techniques is the integration of survey software with widely used statistical analysis systems. A leader in this trend from the early 1980's is the SUDAAN system, which is comprised of a series of several SAS procedures. Freeman *et al.* (1985) and Hidiroglou and Paton (1987) both used the PROC MATRIX procedure in SAS to obtain survey variance estimates, the former by balanced repeated replication and the latter by Taylor linearization. Mohadjer *et al.* (1986) report the development of a new SAS procedure WESVAR to obtain survey variance estimates by balanced repeated replication.

A variety of packages and computing techniques are available to carry out the analyses of survey data reviewed in Section 4. Among the available specialized packages, the most comprehensive appears to be the PC CARP. The original program, SUPER CARP, was designed to carry out regression analyses developed by Fuller (1975); the PC version retains this option. The current version now contains additional options for categorical data analysis, and inferences on cumulative distribution function and associated quantiles, following methods given by Francisco and Fuller (1986). For categorical data, there is an option for the analysis of two-way contingency tables, based on the Rao-Scott corrections to chi-squared test of independence. The program can also be manipulated to perform factor analyses of survey data.

There are four other specialized packages for the analysis of survey data; between them they cover topics in regression and categorical data analysis. The &REPERR program in OSIRIS IV and the SURREGR procedure in SUDAAN both calculate standard errors of regression coefficients so that regression analyses can be carried out. The programs CPLX, developed by Fay (1982), and RSPLX, also by Fay, handle categorical data analyses of log-linear models for two and multi-way tables. The analysis in CPLX is carried out using jackknifed chi-square statistics, while RSPLX applies second order Rao-Scott corrections to the usual test statistic.

The four programs for the regression analyses for complex survey data were evaluated by Cohen, Xanthopoulis and Jones (1988). The older version, SUPER CARP, was included in this analysis rather than PC CARP. Similar to the earlier study of Cohen, Burt and Jones (1986) on variance estimation, data from the National Medical Care Expenditure Survey were used. Once again, a program in the SUDAAN suite of programs, SURREGR, was the most

efficient in terms of CPU time usage and easier to program than the others. However, the efficiency of the SUDAAN programs might be balanced by the flexibility of the PC CARP program, depending upon the survey analysis required.

Significant enhancements to SUDAAN are provided in the new SUDAAN system under development (LaVange *et al.* 1989). Variance estimation and data analysis methods not available in SUDAAN are among the many modifications incorporated into the new SUDAAN System.

Running almost parallel to the emerging trend in the calculation of variance estimates, there is pagea move towards incorporating methods for the analysis of complex survey data into standard statistical packages and systems. Following on their variance estimation methods using SAS procedures, Hidiroglou and Paton (1987) describe further SAS procedures to carry out log-linear analyses, with Rao-Scott corrections, of multi-way contingency tables. Likewise, Freeman (1988) notes that he used the SAS procedure PROC MATRIX for both variance estimation and for the analysis of variance of his survey data. Similarly, Mahodjer *et al.* (1986) describe two other new SAS procedures in addition to the variance estimation procedure WESVAR. These are the previously mentioned NASSREG and NASSLOG which carry out weighted least squares regression analyses and logistic regression analyses respectively. Both procedures depend on balanced repeated replication for variance estimation of the model parameters. An alternative approach to using SAS procedures is to use the matrix algebra language GAUSS (Platt 1986). Based on their own experience, Rao and Thomas (1988) favorably report on the use of this language for categorical data analysis in complex surveys.

6. CONCLUDING REMARKS

The early milestones in the development of efficient sampling designs and associated estimation techniques for population totals and means have firmly established sample survey theory and methods as a major discipline in statistics. Subsequent developments in theoretical foundations of sampling theory have provided useful insights into inferential aspects. In particular, the model-assisted approach and the conditional design-based approach appear to be promising since they attempt to fill the "gap" between the traditional approach and the model-dependent approach by retaining the desirable features of both approaches, but more research is needed in this area to handle complex sampling designs. Recent advances in variance estimation and confidence intervals for nonlinear statistics and the associated computer software, are also equally impressive. It is also gratifying that rapid progress has been made in the development of methods for the analysis of survey data that take account of the complexity of the sampling design, and the associated computer software.

We can expect to see important new developments in the next 10 years or so in the areas of variance estimation for nonlinear statistics (especially, nonsmooth functions), analysis of survey data (especially, multivariate analysis), and other topics not covered here (especially, sampling in time and small area estimation).

ACKNOWLEDGEMENTS

The authors would like to thank the editor for helpful comments. This work was supported by research grants from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

BASU, D. (1971). An essay on the logical foundations of survey sampling, Part I. In *Foundations of Statistical Inference*, (Eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.

BEBBINGTON, A.C., and SMITH, T.M.F. (1977). The effect of survey design on multivariate analysis. In *The Analysis of Survey Data* (Eds. C.A. O'Muircheartaigh and C.D. Payne), Vol. 2, New York: Wiley, 175-192.

BEDRICK, E.J. (1983). Adjusted goodness-of-fit tests for survey data. *Biometrika*, 70, 591-595.

BELLHOUSE, D.R. (1988). A brief history of random sampling methods. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland, 1-14.

BELLHOUSE, D.R., and RAO, J.N.K. (1986). On the efficiency of prediction estimators in two-stage sampling. *Journal of Statistical Planning and Inference*, 13, 269-281.

BINDER, D.A. (1983). On the variance of the asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

BOWLEY, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, Supplement to Liv.1, 6-62.

BREWER, K.R.W. (1963). Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

BREWER, K.R.W., and HANIF, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.

CASSEL, C.M., SÄRNDAL, C.E., and WRETMAN, J.H. (1976). *Foundations of Inference in Survey Sampling*. New York: Wiley.

CHAUDHURI, A. (1988). Optimality of sampling strategies. In *Handbook of Statistics* (Eds. P.R. Krishmaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland, 47-96.

CHAUDHURI, A., and VOS, J.W.E. (1988). *Unified Theory and Strategies of Survey Sampling*. Amsterdam: North-Holland.

COCHRAN, W.G. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.

COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.

COHEN, S.B., BURT, V.L., and JONES, G.K. (1986). Efficiencies in variance estimation for complex survey data. *American Statistician*, 40, 157-164.

COHEN, S.B., XANTHOPOULIS, J.A., and JONES, G.K. (1988). An evaluation of statistical software procedures appropriate for the regression analysis of complex survey data. *Journal of Official Statistics*, 4, 17-34.

DALENIUS, T. (1957). *Sampling in Sweden*. Stockholm: Almquist and Wiksell.

DEMING, W.E. (1960). *Sample Design in Business Research*. New York: Wiley.

DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling* (Eds. N.L. Johnson and H. Smith), New York: Wiley-Interscience, 629-651.

ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society*, Series B, 31, 195-224.

FAY, R.E. (1982). Contingency tables for complex designs, CPLX. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 44-53.

FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

FELLEGI, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, 261-268.

FIELLER, E.C. (1932). The distribution of the index in a normal bivariate population. *Biometrika*, 24, 428-440.

FISHER, R.A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd 5th edition 1934.

FOLSOM, R.E. (1974). National assessment approach to sampling error estimation, sampling error monograph. National Assessment of Educational Progress, first draft.

FRANCISCO, C.A., and FULLER, W.A. (1986). Estimation of the distribution function with a complex survey. Technical Report, Iowe State University.

FREEMAN, D.H. (1988). Sample survey analysis: analysis of variance and contingency tables. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland, 415-426.

FREEMAN, D.H., LIVINGSTON, M., LEO, L., and LEAF, P. (1985). A comparison of indirect variance estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 313-316.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, Series C, 37, 117-132.

FULLER, W.A. (1987). Estimators of the factor model for survey data. In *Applied Probability, Statistics and Sampling Theory* (Eds. I.B. MacNeill and G.J. Umphrey), Boston: D. Reidel Publishing Company, 265-284.

GHOSH, M. (1987). On admissibility and uniform admissibility in finite population sampling. In *Applied Probability, Stochastic Processes and Sampling Theory*, (Eds. I.B. MacNeil and G.J. Umphrey), Boston: D. Reidel Publishing Company, 197-213.

GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society*, series B, 17, 269-278.

GODAMBE, V.P. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society*, series B, 28, 310-328.

GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54,127-138.

GROSS, W.F. (1984). A note on chi-squared tests with survey data. *Journal of the Royal Statistical Society*, series B, 46, 270-272.

GUPTA, V.K., and NIGAM, A.K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika*, 74, 735-742.

HÁJEK, J. (1981). *Sampling From a Finite Population*. New York: Marcel Dekker.

HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sampling Survey Methods and Theory*, Vol. 1. New York: Wiley.

HANSEN, M.H., MADOW, W.G., and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

HANSEN, M.H., DALENIUS, T., and TEPPING, B.J. (1985). The development of sample surveys of finite populations. In *A Celebration of Statistics*: The ISI Centenary Volume (Eds. A.C. Atkinson and S.E. Fienberg), New York: Springer Verlag, 327-354.

HARTLEY, H.O., and RAO, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

HIDIROGLOU, M.A., FULLER, W.A., and HICKMAN, R. (1980). *SUPERCARP–Sixth Edition*. Survey Section, Ames, Iowa.

HIDIROGLOU, M.A., and PATON, D.J. (1987). Some experiences in computing estimates and their variances using data from complex survey designs. In *Applied Probability, Statistics and Sampling Theory* (Eds. I.B. MacNeill and G.J. Umphrey), Boston: D. Reidel Publishing Company, 285-308.

HOLT, D., and SMITH T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society*, Series A, 142, 33-46.

HOLT, M.M. (1979). SURREGR: standard errors of regression coefficients from sampling survey data. Research Triangle Institute, Research Triangle Park, North Carolina.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

ISAKI, C.T. and FULLER, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, No. 304.

JONES, G.K. (1983). HESBRR (HES variance and crosstabulation program). Version 3, Internal NCHS Report, Hyattsville, Maryland.

KIAER, A. (1897). The representative method of statistical surveys (1976 English translation of the original Norwegian), Oslo. Central Bureau of Statistics of Norway.

KISH, L. (1965). *Survey Sampling*. New York: Wiley.

KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of Royal Statistical Society*, series B, 36, 1-37.

KOCH, G.G., FREEMAN, D.H., and FREEMAN, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, 43, 59-78.

KOVAR, J., RAO, J.N.K., and WU, C.F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, Supplement, 25-45.

KOTT, P.S. (1987). Estimating the conditional variance of a design consistent regression estimator. Technical Report.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife, and balanced repeated replication methods. *Annals Statistics*, 9, 1010-1019.

LAVANGE, L.M., SHAH, B.V., BARNWELL, B.G., and KILLINGER, J.F. (1989). SUDAAN: A comprehensive package for survey data analysis. Technical Report, Research Triangle Institute.

LEPKOWSKI, J.M. (1982). The use of OSIRIS IV to analyse complex sample survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 38-43.

LITTLE, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.

MADOW, W.G. (1978). Comments on papers by Basu and Royall and Cumberland. In *Survey Sampling and Measurement* (Ed. N.K. Namboodiri). New York: Academic Press, 315-322.

MADOW, W.G., and MADOW, L.H. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15, 1-24.

MAHALANOBIS, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-370.

McCARTHY, P.J. (1969). Pseudo-replication: half-samples. *International Statistical Review*, 37, 239-264.

MOHADJER, L., MORGANSTEIN, D., CHU, A., and RHOADS, M. (1986). Estimation and analysis of survey data using SAS procedures WESVAR, NASSREG, and NASSLOG. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 258-263.

MONTANARI, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.

NARAIN, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.

NATHAN, G. (1988). Inference based on data from complex sample designs. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland.

NATHAN, G., and HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society*, series B, 42, 377-386.

NEYMAN, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, series B, 12, 241-255.

PFEFFERMAN, D., and HOLMES, D.J. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society*, series A, 148, 268-278.

PFEFFERMAN, D., and SMITH, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *International Statistical Review*, 53, 37-59.

PLATT, W.G. (1986). GAUSS. *American Statistician*, 40, 164-169.

RAJ, D. (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association*, 61, 391-396.

RAO, J.N.K. (1971). Some thoughts on the foundations of survey sampling. *Journal of the Indian Society of Agricultural Statistics*, 23, 69-82.

RAO, J.N.K. (1979). On deriving mean square errors and their non-negative unbiased estimators. *Journal of the Indian Statistical Association*, 17, 125-136.

RAO, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11, 15-31.

RAO, J.N.K. (1987). Analysis of categorical data from sample surveys. In *New Perspectives in Theoretical and Applied Statistics* (Eds. M.L. Puri, J.P. Vilaplana and W. Wertz). New York: Wiley, 45-60.

RAO, J.N.K. (1988). Variance estimation in sample surveys. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland, 427-447.

RAO, J.N.K., and SINGH, M.P. (1973). On the choice of estimator in survey sampling. *Australian Journal of Statistics*, 15,95-104.

RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data, *Annals of Statistics*, 12, 46-60.

RAO, J.N.K., and WU, C.F.J. (1985). Inference from stratified samples: second order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

RAO, J.N.K., and SCOTT, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.

RAO, J.N.K., and WU, C.F.J. (1988). Resampling inference with complex survey data, *Journal of the American Statistical Association*, 83, 231-241.

RAO, J.N.K., and THOMAS D.R. (1988). The analysis of cross-classified categorical data from sample surveys. *Sociology Methodology*, 18, 213-269.

ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

ROBINSON, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association*, 82, 826-831.

ROYALL, R.M. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63, 1269-1279.

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

ROYALL, R.M., and HERSON, J.H. (1973). Robust estimation in finite populations, I and II. *Journal of the American Statistical Association*, 68, 880-889 and 890-893.

ROYALL, R.M., and EBERHARDT, K.R. (1975). Variance estimates for the ratio estimator. *Sankhyā*, series C, 37, 43-52.

ROYALL, R.M., and PFEFFERMAN, D. (1982). Balanced samples and robust Bayesian inference in finite population sampling. *Biometrika*, 69, 401-410.

ROYALL, R.M., and CUMBERLAND, W.G. (1988). Does simple random sampling provide adequate balance? *Journal of the Royal Statistical Society*, series B, 50, 118-124.

RUST, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 4, 381-397.

RYLETT, D.T., and BELLHOUSE, D.R. (1988). TREES: a computer program for complex surveys. *Proceedings of the Section on Survey Research Methods*. American Statistical Association,694-697.

SÄRNDAL, C.E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SÄRNDAL, C.E., and WRIGHT, R.L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.

SÄRNDAL, C.E., SWENSSON, B., and WRETMAN, J.H. (1989). The weighted regression technique for estimating the variance of the generalized regression estimator. *Biometrika*, 76, 527-537.

SCHNELL, D., SULLIVAN, G., KENNEDY, W.J., and FULLER, W.A. (1986). PC CARP: Variance estimation for complex surveys. In *Computer Science and Statistics: Proceedings of the 17th Symposium of the Interface* (Ed. D.M. Allen). Amsterdam: North Holland, 125-129.

SCHNELL, D., KENNEDY, W.J., SULLIVAN, G., PARK, H.J., and FULLER, W.A. (1988). Personal computer variance software for complex surveys. *Survey Methodology*, 14, 59-69.

SCOTT, A.J., and SMITH, T.M.F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 64, 830-840.

SCOTT, A.J., and HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.

SHAH, B.V. (1981a). SESUDAAN: Standard errors program for computing of standardized rates from sample survey data. Research Triangle Institute, Research Triangle Park, North Carolina.

SHAH, B.V. (1981b)., RATIOEST: Standard errors program for computing ratio estimates for sample survey data. Research Triangle Institute, Research Triangle Park, North Carolina.

SHAH, B.V. (1982). RTIFREQS: Program to compute weighted frequencies, percentages and their standard errors. Research Triangle Institute, Research Triangle Park, North Carolina.

SKINNER, C.J., HOLMES, D.J., and SMITH, T.M.F. (1986). The effect of sample design on principal component analysis. *Journal of the American Statistical Association*, 81, 789-798.

SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: Wiley.

SMITH, T.M.F. (1984). Present position and potential developments: some personal views - sample surveys. *Journal of the Royal Statistical Society*, series A, 147, 208-221.

THOMAS, D.R., and RAO, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

TORTORA, R.D. (1980). The effect of disproportionate stratified design on principal component analysis used for variable elimination. *Proceedings of the Survey Research Section*, American Statistical Association, 746-750.

VERMA, V., and PEARCE, M.(1977). Users manual for CLUSTERS: A sampling program for computation of sampling errors for clustered samples. Technical Report No. 568, World Fertility Survey, U.K.

VINTER, S. (1980). Survey sampling errors with OSIRIS IV. COMPSTAT *1980: Proceedings in Computational Statistics*, Vienna: Physica-Verlag, 72-80.

WANG, J.C., and WU, C.F.J. (1988). An approach to the construction of asymmetrical orthogonal arrays. Technical Report, University of Waterloo.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures, *Journal of the American Statistical Association*, 47, 635-646.

WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.

WOODRUFF, R.S., and CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.

WU, C.F.J., HOLT, D., and HOLMES, D.J. (1988). The effect of two-stage sampling on the *F* statistic. *Journal of the American Statistical Association*, 83, 150-159.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, Third Edition, London: Griffin.

COMMENT

T.M. FRED SMITH[1]

Sample surveys are one of the most important areas of the application of statistics. The paper by Professors Rao and Bellhouse is an excellent review of the theoretical development of sample surveys and I find it hard to be critical; but in the best traditions of the Royal Statistical Society I shall make the attempt in as constructive and a controversial manner as possible. In any review paper the choice of topics, especially relating to recent work, must be to some extent subjective. This affords a discussant an easy target; criticize the authors for their sins of omission. Also a review must be wide ranging and this allows discussants freedom to ride their own hobby horses over the range. I shall adopt both approaches and my objective in so doing is to identify some additional issues which I believe are important thus widening the review still further.

There is now general agreement about the milestones of our subject. These are associated with the names of Kiaer, Bowley, Neyman, Cochran, Hansen, Hurwitz, Madow, Mahalanobis, Horvitz and Thompson - an international collection dominated, latterly by contributions from the USA. Kiaer and Bowley's work was fundamental because they demonstrated that valid conclusions could be drawn from representative samples of quite small size drawn from large populations with arbitrary values. Representative samples were stratified samples with proportional allocation, and Bowley derived the appropriate theoretical results. Neyman and subsequent authors argued the case for random sampling and developed a comprehensive theory of randomisation inference applicable to most sampling schemes. Durbin (1953) completes the theory with his multi-stage sampling results. Despite the importance of these results sample surveys became a Cinderella subject on the fringes of mainstream statistics, and even today most university departments do not have a sampling statistician on their staff. Why is this?

One reason is that sample survey theory has developed mainly within social science and official government statistics, whereas most statisticians have a training within mathematics and physical science. Although all experimental scientists deal with samples very few seem to recognise this explicitly and those that do, such as geologists and biologists, have developed their own theory of sampling and estimation. In my view it is time to bring together sampling experts from all areas of scientific enquiry to share ideas and experiences and hopefully to establish a global theory of sample surveys.

A second reason is that sample surveys starts with a population which is a real fixed finite population of units. Samples are then drawn from this population according to specific rules. In most scientific enquiries the position is reversed; the population is not well defined and the scientist starts with a sample. One view of the role of the statistician, as enunciated, for example, R.A. Fisher, is to define the hypothetical population from which the sample data can be viewed as a random sample. This approach begs the question whether this hypothetical population has any scientific value. Arguably the sample survey approach of starting with the population has much to commend it.

A third reason is that since the finite population units can take arbitrary values the population cannot be summarized by a few parameters. Notions like sufficiency have little value in sample survey theory, and sample data are usually summarized by a mass of cross-tabulations. The estimation of a large number of cell proportions is the primary aim of sample surveys and the object of inference is usually descriptive rather than explanatory.

A final reason for the separation of sample surveys from mainstream statistics is that the randomisation theory of sample surveys is so complete. It is a closed theory which if accepted

[1] T.M.F. Smith, Department of Mathematics, The University, Southampton, S09 5NH, U.K.

has few remaining problems to be solved. The chief concerns of randomisation researchers since Horvitz and Thompson (1952) provided the general theoretical framework have been the construction of πps sampling schemes with non-zero joint inclusion probabilities, the production of methods and programs for variance estimation and the construction of estimators which employ auxiliary information but can never be generally efficient because of Godambe's result. All of these problems are important, but they are not exciting, they lack the philosophical and mathematical depth to capture the imaginations of young mathematical statisticians.

These reasons are my explanation why sample surveys have been seen in the past as an activity on the fringe of mainstream statistics. The position is changing now and I detect a coming together of the branches of statistics. Much recent work in sample surveys has attempted to integrate surveys into mainstream statistics and many areas of statistics now recognise the importance of selection effects. Has the sample survey Cinderella been invited to the Statisticians's Ball?

In addition to his non-existence theorem Godambe has also shown that within the randomisation framework the likelihood is proportional to the probability of selection, $p(s \mid z)$, where z is the prior information on which the design was based, which for fixed s is a constant. Thus the likelihood is completely uninformative. In the same set-up Basu (1971) showed that the sufficient statistic is $\{(i,y_i):i\in s\}$, namely the complete data tape including the labels. Although these results are also negative, highlighting the distinction between randomisation inference and other forms of inference, they did stimulate interest amongst a wider group of statistician and so had a positive value. My own interest in the theory of sample surveys was stimulated by Ericson (1969), in particular by the way he incorporated the uninformative likelihood into a positive framework via Bayes theorem and exchangeable priors. Ericson's use of exchangeability deserves consideration by all statisticians, not just Bayesians. Is it reasonable, is it even possible, to have a valid theory of predictive inference without some form of exchangeability? If there is no function of the unit values which is exchangeable how can you predict the unobserved values from the sample values? My opinion is that Ericson's work was a milestone in the development of sample survey theory.

The uninformative nature of the randomisation likelihood led some statisticians to question the role of randomisation. Godambe himself refers to "the problem of randomisation" and developed alternative theoretical approaches which required randomisation. Ericson also found a role for randomisation within his exchangeable set-up. He argued that if you employ your prior information, z, to form groups of units which are approximately exchangeable a priori then the use of simple random sampling will guarantee exchangeability. Royall (1970, 1973), however, made the mistake of advocating purposive sampling within his model-based framework. He touched a raw nerve and brought down upon his head the wrath of the randomisation establishment. I thought that Royall had asked some serious questions which deserved an answer and the strength of the reaction surprised me. Why did academic survey samplers and those from government agencies in North America feel so strongly about randomisation? Their colleagues in market research seemed happy with quota samples which could be viewed as a special case of balanced sampling. In Europe many official surveys are based on quota samples. What is so special about official statistics in North America?

I think the answer lies deep in the American political psyche. Thoughtful Americans are democratic in the true sense of that term. They believe in individual freedom and the right to information, they are also deeply suspicious of governments. They recognise the need within a democracy for reliable statistical information. To the official statisticians randomisation is the guarantee of the objective reliability of their data. It is a key source of their professional integrity and any attack on randomisation was seen as potentially dangerous however well

intentioned. I admire this position and it has helped to convince me that randomisation is one of the great contributions of statistics to science.

I have expressed myself with some feeling because I am so unhappy about the present position of official statistics in the U.K.. The tradition in the U.K. is not naturally democratic, we are still a monarchy, we respect authority rather than the individual. This tendency is being exploited and there is now a serious erosion of public confidence in the Government's use of statistics. It has been argued that official statistics in the U.K. are collected to aid the decisions of government, not to help parliament or to inform the electorate. Key series have been stopped, definitions have been changed, information is presented by ministers in ways which are patently false, yet no government statistician can complain publically because of the Official Secrets Act. There is a dangerous public cynicism about statistics and George Orwell's predictions in his novel 1984 may be closer to the truth than we realise. I apologise to the authors for this digression, but I said I would ride some hobby horses, and the issue of the integrity of official statistics is of great importance.

Before leaving randomisation theory I would like to make some comments about repeated surveys and rotation sampling. Again this is an area which the authors have excluded although they did note Patterson (1950) as a milestone paper. Randomisation theory has been developed within the framework of the one-off cross section survey. The extension to repeated surveys is non-trivial for it is difficult to retain the probability structure over time under rotation sampling when the population changes, Fellegi (1963). For the measurement of gross flows, or transition probabilities, the role of the randomisation inclusion probabilities is not clear. The beautiful simplicity of randomisation theory for one-off surveys is destroyed when they are repeated over time. But most important surveys are repeated surveys, especially in the government sector, so what are the implications?

As always the answer is that it depends. If the primary purpose is to produce descriptive statistics of the state of the system at each time period then the surveys can be considered as repetitions of a cross-section survey and each one can be analysed independently. Although composite estimators or time series estimators may be more efficient they should be viewed as secondary estimators rather than primary estimators. If I wanted to use repeated survey data within an econometric model I would prefer to input the cross-section estimates with their known correlation structure rather than complex composite estimates. On the other hand if I wanted the best estimate of the current value of, say, unemployment, for a particular purpose, not for public consumption, then I would use the most efficient procedure available. Similarly if I wanted to explain the change in value of some estimates over time then I would need to go beyond simple randomisation analysis. Thus the problems with randomisation inference for repeated surveys occur mainly for secondary analyses. However, there remains the important issue of which estimates should be reported to the public.

Section 2 of the paper is devoted to work on the theoretical foundations of inference from survey data carried out during the last 30-40 years. The authors have chosen to distinguish three approaches, design-based, model-dependent and model-assisted, the latter being an attempt to find a compromise solution between the other two. Personally I prefer to go for a GUT (Grand Universal Theory) approach integrating both design and models into one framework. The important influences on my thinking in this area, in addition to Ericson, have been Scott (1977) and Rubin (1976). In the GUT approach the survey variables, the sampling mechanism, and any other selection and measurement mechanisms are all introduced explicitly into an overall model. If Y is the $n \times p$ matrix of measured survey variables, z is the prior information, s denotes the sample, $s^* \subset s$ denotes the respondents, then the joint distribution of all these variables is

$$f(Y \mid z;\Theta)g(z;\phi)\mathrm{p}(\mathrm{s} \mid z)\,q(s^* \mid s,z,Y_s;\eta),$$

where the survey design, represented by $p(s \mid z)$, is of the so-called uninformative type such as random sampling. The design is uninformative because z is assumed known and includes all the usual information on stratification, clustering and measures of size. This general formulation forces statisticians to face up to all their assumptions. Non-response must be modelled explicitly. Measurement errors must be included in the structure of $f(Y \mid z;\Theta)g(z;\phi)$. The decision to use randomisation inference is then an explicit statement that given z the values of Y can be treated as unknown constants; they are arbitrary values about which we have no additional information. A modeller, on the other hand must specify the model to the level needed for inference, for example, by an exchangeable model. Both design-based and model-dependent approaches condition on the same prior information, z, and so both should employ similar, possibly identical, structures. In fact I would rarely expect the point estimators using the different approaches to differ very much in practice. The issue thus becomes that identified by the authors as the choice of a measure of uncertainty. Model-dependent procedures employ conditional variances, strict design-based procedures are unconditional. How to construct conditional design-based inferences is still an open question, but the approach of Robinson (1987) looks promising. The GUT model shows the design-based versus model-based controversy to be what it is, namely a relatively small philosophical dispute within the much bigger framework of total survey analysis.

The failure of both theoretical and practical statisticians to integrate sampling and non-sampling errors into measures of total survey error even after 50 years of intensive research must be noted as one of the failures of this important branch of statistics. But again things are changing and the mood now is no longer merely to report sampling errors and in addition to give vague warnings about the potential size of non-sampling errors but it is to attempt to measure total survey error recognising that some non-sampling biases can far exceed sampling errors.

Section 4 of the paper is devoted to the analysis of survey data, to the analytic rather than descriptive uses of surveys. Here the design-based, model-based dispute pales into insignificance. Analysts must face up to all the classical problems of model choice, estimation and testing, residual analysis and so on, which make up mainstream statistics. Cinderella is at last dancing with the Prince.

My final comments are again personal. If you look at the references at the end of the paper, and if you consider the additional areas which I have discussed, then you will see that Jon Rao has contributed important papers in every area. I think that it was particularly appropriate that he was invited to write this paper. I congratulate both authors on their fine paper.

ADDITIONAL REFERENCES

DURBIN, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. *Journal of the Royal Statistical Society, Series B, 15*, 262-269.

ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B, 31*, 195-233.

FELLEGI, I.P. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

SCOTT, A.J. (1977). On the problem of randomization in survey sampling. *Sankhyā*, C, 39, 1-9.