Interplay Between Sample Survey Theory and Practice: an Appraisal

J. N. K. Rao

ABSTRACT

A large part of sample survey theory has been directly motivated by practical problems encountered in the design and analysis of sample surveys. On the other hand, sample survey theory has influenced practice, often leading to significant improvements. This paper will examine this interplay over the past 60 years or so. Examples where new theory is needed or where theory exists but is not used will also be presented.

KEY WORDS: Analysis of survey data; Early contributions; Inferential issues; Re-sampling methods; Small area estimation.

1. SOME EARLY LANDMARK CONTRIBUTIONS: 1920-1970

This section gives an account of some early landmark contributions to sample survey theory and methods that have greatly influenced the practice. The Norwegian statistician A. N. Kier (1897) is perhaps the first to promote sampling (or what was then called "the representative method") over complete enumeration, although the oldest reference to sampling can be traced back to the great Indian epic Mahabharata (Hacking, 1975, p.7). In the representative method the sample should mirror the parent finite population and this may be achieved either by balanced sampling through purposive selection or by random sampling. The representative method was used in Russia as early as 1900 (Zarkovic, 1956) and Wright conducted sample surveys in the United States around the same period using this method. By the 1920's, the representative method was widely used, and the International Statistical Institute played a prominent role by creating a committee in 1924 to report on the representative method. This committee's 1925 report discussed theoretical and practical aspects of the random sampling method. Bowley's (1926) contribution to this report includes his fundamental work on stratified random sampling with proportional allocation, leading to a representative sample with equal inclusion probabilities. Hubback (1927) recognized the need for random sampling in crop surveys: "The only way in which a satisfactory estimate can be found is by as close an approximation to random sampling as the circumstances permit, since that not only gets rid of the personal limitations of the experimenter but also makes it possible to say what is the probability with which the results of a given number of samples will be within a given range from the mean. To put this into definite language, it should be possible to find out how many samples will be required to secure that the odds are at least 20:1 on the mean of the samples within one maund of the true mean". This statement contains two important observations on random sampling: (1). It avoids personal biases in sample selection. (2). Sample size can be determined to satisfy a specified margin of

error apart from a chance of 1 in 20. Mahalanobis (1946b) remarked that R. A. Fisher's fundamental work at Rothamsted Experimental Station on design of experiments was influenced directly by Hubback (1927).

Neyman's (1934) classic landmark paper laid the theoretical foundations to the probability sampling (or design-based) approach to inference from survey samples. He showed, both theoretically and with practical examples, that stratified random sampling is preferable to balanced sampling because the latter can perform poorly if the underlying model assumptions are violated. Neyman also introduced the ideas of efficiency and optimal allocation in his theory of stratified random sampling without replacement by relaxing the condition of equal inclusion probabilities. By generalizing the Markov theorem on least squares estimation, Neyman proved that the stratified mean , $\bar{y}_{st} = \sum_h W_h \bar{y}_h$, is the best estimator of the population mean, $\bar{Y} = \sum_h W_h \bar{Y}_h$, in the linear class of unbiased estimators of the form $\bar{y}_b = \sum_h W_h \sum_i b_{hi} y_{hi}$, where $W_h, \bar{y}_h$ and $\bar{Y}_h$ are the $h-th$ stratum weight, sample mean and population mean ($h = 1,...,L$), and b_{hi} is a constant associated with the item value y'_{hi} observed on the $i-th$ sample draw ($i = 1,...,n_h$) in the $h-th$ stratum. Optimal allocation $(n_1,...,n_L)$ of the total sample size, n, was obtained by minimizing the variance of $\bar{y}_{st}$ subject to $\sum_h n_h = n$; an earlier proof of Neyman allocation by Tschuprow (1923) was later discovered. Neyman also proposed inference from larger samples based on normal theory confidence intervals such that the frequency of errors in the confidence statements based on all possible stratified random samples that could be drawn does not exceed the limit prescribed in advance *"whatever the unknown properties of the population"*. Any method of sampling that satisfies the above frequency statement was called "representative". Note that Huback (1927) earlier alluded to the frequency statement associated with the confidence interval. Neyman's final contribution to the theory of sample surveys (Neyman, 1938) studied two-phase sampling for stratification and derived the optimal first phase and second phase sample sizes, n' and n, by minimizing the variance of the estimator subject to a given cost $C = n'c' + nc,$ where the second phase cost per unit, $c,$ is large relative to the first phase cost per unit, c'.

The 1930's saw a rapid growth in demand for information, and the advantages of probability sampling in terms of greater scope, reduced cost, greater speed and model-free features were soon recognized, leading to an increase in the number and type of surveys taken by probability sampling and covering large populations. Neyman's approach was almost universally accepted by practicing survey statisticians. Moreover, it inspired various important extensions, mostly motivated by practical and efficiency considerations. Cochran's (1939) landmark paper contains several important results: the use of ANOVA to estimate the gain in efficiency due to stratification, estimation of variance components in two-stage sampling for future studies on similar material, choice of sampling unit, regression estimation under two-phase sampling and effect of errors in strata sizes. This paper also introduced the super-population concept: "The finite

population should itself be regarded as a random sample from some infinite population". It is interesting to note that Cochran at that time was critical of the traditional fixed population concept: " Further, it is far removed from reality to regard the population as a fixed batch of known numbers". Cochran (1940) introduced ratio estimation for sample surveys, although an early use of the ratio estimator dates back to Laplace (1820). In another landmark paper (Cochran, 1942), he developed the theory of regression estimation. He derived the conditional variance of the usual regression estimator for a fixed sample and also a sample estimator of this variance, assuming a linear regression model $y = \alpha + \beta x + e$, where e has mean zero and constant variance in arrays in which x is fixed. He also noted that the regression estimator remains (model) unbiased under non-random sampling, provided the assumed linear regression model is correct. He derived the average bias under model deviations (in particular, quadratic regression) for simple random sampling as the sample size n increased. Cochran then extended his results to weighted regression and derived the now well-known optimality result for the ratio estimator, namely it is a " best unbiased linear estimate if the mean value and variance both change proportional to x". The latter model is called the ratio model in the current literature. Cochran (1946) compared the expected (or anticipated) variance under a super-population model to study the relative efficiency of alternative probability sampling strategies (design and estimator) analytically. This paper stimulated much subsequent research on the use of super-population models in the choice of probability sampling strategies, and also for model-dependent and model- assisted inferences (see Section 2).

In India, Mahalanobis made pioneering contributions to sampling by formulating cost and variance functions for the design of surveys. His 1944 landmark paper (Mahalanobis, 1944) provides deep theoretical results on the efficient design of sample surveys and their practical applications, in particular to crop acreage and yield surveys. The well-known optimal allocation in stratified random sampling with cost per unit varying across strata is obtained as a special case of his general theory. As early as 1937, Mahalanobis used multi-stage designs for crop yield surveys with villages, grids within villages, plots within grids and cuts of different sizes and shapes as sampling units in the four stages of sampling (Murthy, 1964). He also used a two-phase sampling design for estimating the yield of cinchona bark. He was instrumental in establishing the National Sample Survey (NSS) of India, the largest multi-subject continuing survey operation with full-time staff using personal interviews for socioeconomic surveys and physical measurements for crop surveys. Several eminent survey statisticians, including D.B. Lahiri and M. N. Murthy, were associated with the NSS.

P. V. Sukhatme, who studied under Neyman, also made pioneering contributions to the design and analysis of large-scale agricultural surveys in India, using stratified multi-stage sampling. Begining in 1942-43 he developed efficient designs for the conduct of nationwide surveys on wheat and rice crops and demonstrated high degree of precision for state estimates and reasonable margin of error for district estimates. Sukhatme's approach differed from that of Mahalanobis who used very small plots for crop cutting employing *ad hoc* staff of investigators. Sukhatme (1947) and Sukhatme and Panse

(1951) showed that the use of a small plot might give biased estimates due to the tendency of placing boundary plants inside the plot when there is doubt. They also pointed out that the use of *ad hoc* staff of investigators, moving rapidly from place to place, forces the plot measurements on only those sample fields that are ready for harvest on the date of the visit, thus violating the principle of random sampling. Sukhatme's solution was to use large plots to avoid boundary bias and to entrust crop-cutting work to the local revenue or agricultural agency in a State.

Survey statisticians at the U. S. Census Bureau, under the leadership of Morris Hansen, William Hurwitz, William Madow and Joseph Waksberg, made fundamental contributions to sample survey theory and practice during the period 1940-70, and many of those methods are still widely used in practice. Hansen and Hurwitz (1943) developed the basic theory of stratified two-stage sampling with one primary sampling unit (PSU) within each stratum drawn with probability proportional to size measure (PPS sampling) and then sub-sampled at a rate that ensures self-weighting (equal overall probabilities of selection) within strata. This approach provides approximately equal interviewer work loads which is desirable in terms of field operations. It also leads to significant variance reduction by controlling the variability arising from unequal PSU sizes without actually stratifying by size and thus allowing stratification on other variables to reduce the variance. On the other hand, workloads can vary widely if the PSUs are selected by simple random sampling and then sub-sampled at the same rate within each stratum. PPS sampling of PSUs is now widely used in the design of large-scale surveys, but two or more PSUs are selected without replacement from each stratum such that the PSU inclusion probabilities are proportional to size measures (see Section 4).

Many large-scale surveys are repeated over time, such as the monthly Canadian Labour Force Survey (LFS) and the U. S. Current Population Survey (CPS), with partial replacement of ultimate units (also called rotation sampling). For example, in the LFS the sample of households is divided into six rotation groups (panels) and a rotation group remains in the sample for six consecutive months and then drops out of the sample, thus giving five-sixth overlap between two consecutive months. Yates (1949) and Patterson (1950), following the initial work of Jessen (1942) for sampling on two occasions with partial replacement of units, provided the theoretical foundations for design and estimation of repeated surveys, and demonstrated the efficiency gains for level and change estimation by taking advantage of past data. Hansen et al. (1955) developed simpler estimators, called K-composite estimators, applicable to stratified multi-stage designs with PPS sampling in the first stage. Rao and Graham (1964) studied optimal replacement policies for the K-composite estimators. Various extensions have also been proposed. Composite estimators have been used in the CPS and other continuing large scale surveys. Only recently, the Canadian LFS adopted composite estimation, called regression composite estimation, that makes use of sample information from previous months and that can be implemented with a regression weights software (see Section 3). Keyfitz (1951) proposed an ingenious method of switching to better size measures in continuing surveys based on the latest census counts. His method ensures that the

probability of overlap with the previous sample of PSUs is maximized, thus reducing the field costs and at the same time achieving increased efficiency by using the better size measures in PPS sampling. The Canadian LFS and other continuing surveys have used the Keyfitz method.

The focus of research prior to 1950 was on estimating population totals and means for the whole population and large planned sub-populations, such as states or provinces. However, the users are also interested in totals and means for unplanned sub-populations (also called domains) such as age-sex groups within a province, and parameters other than totals and means such as median and other quantiles, for example median income. Hartley (1959) developed a simple, unified theory for domain estimation applicable to any design, requiring only the standard formulae for the estimator of total and its variance estimator, denoted in the operator notation as $\hat{Y}(y)$ and $v(y)$ respectively. He introduced two synthetic variables ${}_j y_i$ and ${}_j a_i$ which take the values y_i and 1 respectively if the unit i belongs to domain j and equal to 0 otherwise. The estimators of domain total ${}_jY = Y({}_j y)$ and domain size ${}_j N = Y({}_j a)$ are then simply obtained from the formulae for $\hat{Y}(y)$ and $v(y)$ by replacing y_i by ${}_j y_i$ and ${}_j a_i$ respectively. Similarly, estimators of domain means and domain differences and their variance estimators are obtained from the basic formulae for $\hat{Y}(y)$ and $v(y)$. Durbin (1968) also obtained similar results. Domain estimation is now routinely done, using Hartley's ingenious method.

For inference on quantiles, Woodruff (1952) proposed a simple and ingenious method of getting a $(1-\alpha)$ -level confidence interval under general sampling designs, using only the estimated distribution function and its standard error (see Lohr's(1999) book, pp.311-3). Note that the latter are simply obtained from the formulae for a total by changing y to an indicator variable. By equating the Woodruff interval to a normal theory interval on the quantile, a simple formula for the standard error of the $p-th$ quantile estimator may also be obtained as half the length of the interval divided by the upper $\frac{\alpha}{2}$-point of standard N(0,1) distribution which equals 1.96 if α=0.05 (Rao and Wu, 1987; Francisco and Fuller, 1991). A surprising property of the Woodruff interval is that it performs well even when p is small or large and sample size is moderate (Sitter and Wu, 2001).

The importance of measurement errors was realized as early as 1940's. Mahalanobis' (1946a) influential paper developed the technique of interpenetrating sub-samples (called replicated sampling by Deming, 1960). This method was extensively used in large-scale sample surveys in India for assessing both sampling and measurement errors. The sample is drawn in the form of two or more independent sub-samples according to the same sampling design such that each sub-sample provides a valid estimate of the total or mean. The sub-samples are assigned to different interviewers (or teams) which leads to a valid estimate of the total variance that takes proper account of the correlated response

variance component due to interviewers. Interpenetrating sub-samples increase travel costs of interviewers, but they can be reduced through modifictions of interviewer assignments. Hansen et al. (1951) and Sukhatme and Seth (1952) developed basic theories under additive measurement error models, and decomposed the total variance into sampling variance, simple response variance and correlated response variance. The correlated response variance due to interviewers was shown to be of the order k^{-1} regardless of the sample size, where k is the number of interviewers. As a result, it can dominate the total variance if k is not large. The 1950 U.S. Census interviewer variance study showed that this component was indeed large for small areas. As a result, self-enumeration was introduced in the 1960 U.S. Census to reduce this component of the variance. This is indeed a success story of theory influencing practice.

Yet another early milestone in sample survey methods is the concept of design effect (DEFF) due to Leslie Kish (see Kish, 1965, section 8.2). The design effect is defined as the ratio of the actual variance of a statistic under the specified design to the variance that would be obtained under simple random sampling of the same size. This concept is especially useful in the presentation and modeling of sampling errors, and also in the analysis of complex survey data involving clustering and unequal probabilities of selection (see Section 5).

2. INFERENTIAL ISSUES

2.1 Unified design-based framework

The development of early sampling theory progressed more or less inductively, although Neyman (1934) studied best linear unbiased estimation for stratified random sampling. Strategies (design and estimation) that appeared reasonable were entertained and relative properties were carefully studied by analytical and /or empirical methods, mainly through comparisons of mean squared errors, and sometimes also by comparing anticipated mean squared errors or variances under plausible super-population models, as noted in Section 1. Unbiased estimation under a given design was not insisted upon because it "often results in much larger mean squared error than necessary" (Hansen et al 1983). Instead, design consistency was deemed necessary for large samples. Classical text books by Cochran (1953), Hansen et al (1953), Sukhatme (1954) and Yates (1949), based on the above approach, greatly influenced survey practice. Yet, academic statisticians paid little attention to traditional sampling theory, possibly because it lacked a formal theoretical framework and not integrated with main stream statistical theory. Graduate courses in sampling theory were not even offered by several prestigious statistics departments in North America.

Formal theoretical frameworks and approaches to integrating sampling theory with main stream statistical inference were initiated in the 1950s under a somewhat idealistic set-up that focussed on sampling errors assuming the absence of measurement or response errors and non-response. Horvitz and Thompson (1952) made a basic contribution to sampling

with arbitrary probabilities of selection by formulating three subclasses of linear design-unbiased estimators of a total Y that include the Markov class studied by Neyman as one of the subclasses. Another subclass with design weight d_i attached to a sample unit i and depending only on i admitted the well-known estimator with weight inversely proportional to the inclusion probability π_i as the only unbiased estimator. Narain (1951) also discovered this estimator, so it should be called Narain-Horvitz-Thompson (NHT) estimator rather than HT estimator as commonly known. For simple random sampling, the sample mean is the best linear unbiased estimator (BLUE) of the population mean in the three subclasses, but this is not sufficient to claim that the sample mean is the best in the class of all possible linear unbiased estimators. Godambe (1955) proposed a general class of linear unbiased estimators of a total Y by recognizing the sample data as $\{(i, y_i), i \in s\}$ and by letting the weight depend on the sample unit i as well as on the other units in the sample s, that is, the weight is of the form $d_i(s)$. He then established that the BLUE does not exist in the general class

$$\hat{Y} = \sum_{i \in s} d_i(s) y_i , \qquad (1)$$

even under simple random sampling. This important negative theoretical result was largely overlooked for about 10 years. Godambe also established a positive result by relating y to a size measure x using a super-population regression model through origin with error variance proportional to x^2, and then showing that the NHT estimator under any fixed sample size design with π_i proportional to x_i minimizes the anticipated variance in the unbiased class (1). This result clearly shows the conditions on the design for the use of NHT estimator but unfortunately some theoretical criteria were later advanced in the sampling literature to claim that the NHT estimator should be used for any sampling design. Rao (1966) recognized the limitations of the NHT estimator in the context of surveys with PPS sampling and multiple characteristics. Here the NHT estimator will be very inefficient when a characteristic y is unrelated or weakly related to the size measure x (such as poultry count y and farm size x in a farm survey). Rao proposed efficient alternative estimators for such cases that ignore the NHT weights. In a well-known example of circus elephants, Basu (1970) constructed a 'bad' design with y_i unrelated to π_i and demonstrated that the NHT estimator leads to absurd estimates which prompted the famous main stream Bayesian statistician Dennis Lindley to even conclude that this counterexample destroys the design-based sample survey theory (Lindley, 1996). This is rather unfortunate because NHT and Godambe clearly stated the conditions on the design for a proper use of the NHT estimator, and Rao (1966) and Hajek (1971) proposed alternative estimators to deal with multiple characteristics and bad designs, respectively. Moreover, a 'good' design in the Basu example would have taken advantage of the past census weights of elephants x which are highly correlated with the current weights y.

Attempts were also made to integrate sample survey theory with mainstream statistical inference via likelihood function. Godambe (1966) showed that the likelihood function from the sample data $\{(i, y_i), i \in s\}$, regarding the N – vector of unknown y – values as the parameter, provides no information on the unobserved sample values and hence on the total Y. This uninformative feature of the likelihood function is due to the label property that treats the N population units as essentially N post-strata. A way out of this difficulty is to ignore some aspects of the sample data to make the sample non-unique and thus arrive at an informative likelihood function (Hartley and Rao, 1968; Royall, 1968). For example, under simple random sampling, suppressing the labels i and regarding the data as $\{(i, y_i), i \in s\}$ in the absence of information relating i to y_i, leads to the sample mean as the maximum likelihood estimator of the population mean. Note that y_i may be a vector that includes auxiliary variables with known totals. In the latter case, Hartley and Rao (1968) showed that the maximum likelihood estimator under simple random sampling is approximately equal to the traditional regression estimator of the total. This paper was the first to show how to incorporate known auxiliary population totals in a likelihood framework. For stratified random sampling, labels within strata are ignored but not strata labels because of known strata differences. The resulting maximum likelihood estimator is approximately equal to a pseudo-optimal linear regression estimator when auxiliary variables with known totals are available. The latter estimator has some good conditional design-based properties (see Section 2.4). The focus of Hartley and Rao (1968) was on point estimation of a total, but the likelihood approach in fact has much wider scope in sampling, including the estimation of distribution functions and quantiles and the construction of likelihood ratio based confidence intervals (see Section 7.1). The Hartley-Rao non-parametric likelihood approach was discovered independently twenty years later (Owen, 1988) in the main stream statistical inference under the name “empirical likelihood” which has attracted a good deal of attention, including its application to various sampling problems. So in a sense the integration efforts with main stream statistics were partially successful. Owen’s (2002) book presents a thorough account of empirical likelihood theory and its applications.

2.2 Model-dependent approach

The model-dependent approach to inference assumes that the population structure obeys a specified super-population model. The distribution induced by the assumed model provides inferences referring to the particular sample of units s that has been drawn. Such conditional inferences are more relevant and appealing than repeated sampling inferences. But model-dependent strategies can perform poorly in large samples when the model is not correctly specified; even small deviations from the assumed model that are not easily detectable through model checking methods can cause serious problems. For example, consider the often-used ratio model when an auxiliary variable x with known total X is also measured in the sample:

$$y_i = \beta x_i + \varepsilon_i ; i = 1,..., N \qquad (2)$$

where the ε_i are independent random variables with zero mean and variance proportional to x_i. Assuming the model holds for the sample, that is, no sample selection bias, the best linear model-unbiased predictor of the total Y is given by the ratio estimator $(\bar{y} / \bar{x})X$ regardless of the sample design. This estimator is not design consistent unless the design is self-weighting, for example, stratified random sampling with proportional allocation. As a result, it can perform very poorly in large samples under non-self-weighting designs even if the deviations from the model are small. Hansen et al. (1983) demonstrated the poor performance under repeated sampling set-up, using a stratified random sampling design with near optimal sample allocation (commonly used to handle highly skewed populations). Rao (1996) used the same design to demonstrate poor performance under a conditional framework relevant to the model-dependent approach (Royall and Cumberland, 1981). Nevertheless, model-dependent approaches can play a vital role in small area estimation where the sample size in a small area (or domain) can be very small or even zero, see Section 6.

Brewer (1963) first proposed the model-dependent approach in the context of the ratio model (2). Royall (1970) and his collaborators made a systematic study of this approach. Valliant, Dorfman and Royall (2000) give a comprehensive account of the theory, including estimation of the (conditional) model variance of the estimator which varies with s, for example, under the ratio model (2) the model variance depends on the sample mean $\bar{x}_s$. It is interesting to note that balanced sampling through purposive selection appears in the model-dependent approach in the context of protection again incorrect specification of the model (Royall and Herson, 1973).

2.3 Model-assisted approach

Model-assisted approach attempts to combine the desirable features of design-based and model-dependent methods. It entertains only design-consistent estimators of the total Y that are also model unbiased under the assumed "working" model. For example, under the ratio model (2), a model-assisted estimator of Y for a specified probability sampling design is given by the ratio estimator $\hat{Y}_r = (\hat{Y}_{NHT} / \hat{X}_{NHT})X$ which is design consistent regardless of the assumed model. Hansen et al. (1983) used this estimator for their stratified design to demonstrate its superior performance over the model dependent estimator $(\bar{y} / \bar{x})X$. For variance estimation, the model-assisted approach uses estimators that are consistent for the design variance of the estimator and at the same time model unbiased for the model variance. However, at the end the inferences are design-based because the model used is only a "working" model and may not hold.

For the ratio estimator $\hat{Y}_r$ the variance estimator is given by

$$s^2(\hat{Y}_r) = (X / \hat{X}_{NHT})^2 v(e), \qquad (3)$$

where in the operator notation $v(e)$ is obtained from $v(y)$ by changing y_i to the residuals $e_i = y_i - (\hat{Y}_{NHT} / \hat{X}_{NHT})x_i$. This variance estimator is asymptotically equivalent to a customary linearization variance estimator $v(e)$, but it reflects the fact that the information in the sample varies with $\hat{X}_{NHT}$: larger values lead to smaller variability and smaller values to larger variability. The resulting normal pivotal leads to valid model-dependent inferences under the assumed model (unlike the use of $v(e)$ in the pivotal) and at the same time protects against model deviations in the sense of providing asymptotically valid design-based inferences. Note that the pivotal is asymptotically equivalent to $\hat{Y}(\tilde{e})/[v(\tilde{e})]^{1/2}$ with $\tilde{e}_i = y_i - (Y / X)x_i$. If the deviations from the model are not large, then the skewness in the residuals $\tilde{e}_i$ will be small even if y_i and x_i are highly skewed, and normal confidence intervals will perform well. On the other hand, in the latter case the normal intervals based on $\hat{Y}_{NHT}$ and its standard error may perform poorly under repeated sampling even for fairly large samples because the pivotal depends on the skewness of the y_i. Therefore, the population structure does matter in design-based inferences contrary to the claims of Neyman (1934), Hansen et al. (1983) and others. Rao et al. (2003) considered the simple linear regression estimator under two-phase simple random sampling with x only observed in the first phase. They demonstrated that the coverage performance of the associated normal intervals can be poor even for moderately large second phase samples if the true underlying model that generated the population deviated significantly from the linear regression model (for example, a quadratic regression of y on x) and the skewness of x is large. In this case, since the first phase x-values will be observed, a proper model-assisted approach would use a multiple linear regression estimator with x and $z = x^2$ as the auxiliary variables. Note that for single phase sampling such a model-assisted estimator cannot be implemented if only the total X is known since the estimator depends on the population total of z.

Sarndal et al. (1992) provide a comprehensive account of the model-assisted apporach to estimating the total Y of a variable y under the working linear regression model

$$y_i = x_i'\beta + \varepsilon_i;\ i = 1,..., N \qquad (4)$$

with mean zero, uncorrelated errors ε_i and model variance $V_m(\varepsilon_i) = \sigma^2 q_i = \sigma_i^2$ where the q_i are known constants and the x-vectors have known totals X (the population values $x_1,..., x_N$ may not be known). Under this set-up, the model-assisted approach leads to the generalized regression (GREG) estimator

$$\hat{Y}_{gr} = \hat{Y}_{NHT} + \hat{B}'(X - \hat{X}_{NHT}) =: \sum_{i \in s} w_i(s) y_i, \quad (5)$$

where

$$\hat{B} = \hat{T}^{-1}(\textstyle\sum_s \pi_i^{-1} x_i y_i / q_i) \quad (6)$$

with $\hat{T} = \sum_s \pi_i^{-1} x_i x_i' / q_i$ is a weighted regression coefficient, and $w_i(s) = g_i(s)\pi_i^{-1}$ with $g_i(s) = 1 + (X - \hat{X}_{NHT})'\hat{T}^{-1} x_i / q_i$, known as "g-weights". Note that the GREG estimator (5) can also be written as $\sum_{i \in U} \hat{y}_i + \hat{E}_{NHT}$, where $\hat{y}_i = x_i'\hat{B}$ is the predictor of y_i under the working model and $\hat{E}_{NHT}$ is the NHT estimator of the total prediction error $E = \sum_{i \in U} e_i$ with $e_i = y_i - \hat{y}_i$. This representation shows the role of the working model in the model-assisted approach. The GREG estimator (5) is design-consistent as well as model-unbiased under the working model (4). Moreover, it is nearly "optimal" in the sense of minimizing the anticipated MSE (model expectation of the design MSE) under the working model provided the inclusion probability, π_i, proportional to the model standard deviation σ_i. However, in surveys with multiple variables of interest, the model variance may vary across variables and one must use a general-purpose design such as the design with inclusion probabilities proportional to sizes. In such cases, the optimality result no longer holds even if the same vector x_i is used for all the variables y_i in the working model.

The GREG estimator simplifies to the 'projection' estimator $X'\hat{B} = \sum_s w_i(s) y_i$ with $g_i(s) = X'\hat{T}^{-1} x_i / q_i$ if the model variance σ_i^2 is proportional to $\lambda' x_i$ for some λ. The ratio estimator is obtained as a special case of the projection estimator by letting $q_i = x_i$, leading to $g_i(s) = X / \hat{X}_{HT}$. Note that the GREG estimator (5) requires only the population totals X and not necessarily the individual population values x_i. This is very useful because the auxiliary population totals are often ascertained from external sources such as demographic projections of age and sex counts. Also, it ensures consistency with the known totals X in the sense of $\sum_s w_i(s) x_i = X$. Because of this property, GREG is also a calibration estimator.

Suppose there are p variables of interest, say $y^{(1)},..., y^{(p)}$, and we want to use the model-assisted approach to estimate the corresponding population totals $Y^{(1)},..., Y^{(p)}$. Also, suppose that the working model for $y^{(j)}$ is of the form (4) but requires possibly different x-vector $x^{(j)}$ with known total $X^{(j)}$ for each $j = 1,..., p$:

$$y_i^{(j)} = x_i^{(j)'} \beta^{(j)} + \varepsilon_i^{(j)}, i = 1,..., N \qquad (7)$$

In this case, the $g-$weights depend on j and in turn the final weights $w_i(s)$ also depend on j. In practice, it is often desirable to use a single set of final weights for all the p variables to ensure internal consistency of figures when aggregated over different variables. This property can be achieved only by enlarging the $x-$vector in the model (7) to accommodate all the variables $y^{(j)}$, say $\tilde{x}$ with known total X and then using the working model

$$y_i^{(j)} = \tilde{x}_i' \beta^{(j)} + \varepsilon_i^{(j)} \ , \ i = 1,..., N \qquad (8)$$

However, the resulting weighted regression coefficients could become unstable due to possible multicolinearity in the enlarged set of auxiliary variables. As a result, the GREG estimator of $Y^{(j)}$ under model (8) is less efficient compared to the GREG estimator under model (7). Moreover, some of the resulting final weights, say $\tilde{w}_i(s)$, may not satisfy range restrictions by taking either values smaller than 1 (including negative values) or very large positive values. A possible solution to handle this problem is to use a generalized ridge regression estimator of $Y^{(j)}$ that is model-assisted under the enlarged model (Chambers, 1996; Rao and Singh, 1997).

For variance estimation, the model-assisted approach attempts to used design-consistent variance estimators that are also model-unbiased (at least for large samples) for the conditional model variance of the GREG estimator. Denoting the variance estimator of the NHT estimator of Y as $v(y)$ in an operator notation, a simple Taylor linearization variance estimator satisfying the above property is given by $v(ge)$, where $v(ge)$ is obtained by changing y_i to $g_i(s)e_i$ in the formula for $v(y)$; see Sarndal *et al.* (1989).

In the above discussion, we have assumed a working linear regression model for all the variables $y^{(j)}$. But in practice a linear regression model may not provide a good fit for some of the $y-$variables of interest, for example, a binary variable. In the latter case, logistic regression provides a suitable working model. A general working model that covers logistic regression is of the form $E_m(y_i) = h(x_i'\beta) = \mu_i$, where $h(.)$ could be non-linear; model (5) is a special case with $h(a) = a$. A model-assisted estimator of the total under the general working model is the difference estimator $\hat{Y}_{NHT} + \sum_U \hat{\mu}_i - \sum_s \pi_i^{-1} \hat{\mu}_i$, where $\hat{\mu}_i = h(x_i'\hat{\beta})$ and $\hat{\beta}$ is an estimator of the model parameter β. It reduces to the GREG estimator (5) if $h(a) = a$. This difference estimator is nearly optimal if the inclusion probability π_i is proportional to σ_i, where σ_i^2 denotes the model variance, $V_m(y_i)$.

GREG estimators have become popular among users because many of the commonly used estimators may be obtained as special cases of (5) by suitable specifications of x_i and q_i. A Generalized Estimation System (GES) based on GREG has been developed at Statistics Canada.

Kott (2005) has proposed an alternative paradigm to inference, called randomization-assisted model-based approach, which attempts to focus on model-based inference assisted by randomization (or repeated sampling). The definition of anticipated variance is reversed to randomization-expected model variance of an estimator, but it is identical to the customary anticipated variance when the working model holds for the sample, as assumed in the paper. As a result, the choices of estimator and variance estimator are often similar to those under the model-assisted approach. However, Kott argues that the motivation is clearer and " the approach proposed here for variance estimation leads to logically coherent treatment of finite population and small-sample adjustments when needed".

2.4 Conditional design-based approach

A conditional design-based approach has also been proposed. This approach attempts to combine the conditional features of the model-dependent approach with the model-free features of the design-based approach. It allows us to restrict the reference set of samples to a "relevant" subset of all possible samples specified by the design. Conditionally valid inferences are obtained in the sense that the conditional bias ratio (i.e., the ratio of conditional bias to conditional standard error) goes to zero as the sample size increases. Approximately $100(1-\alpha)\%$ of the realized confidence intervals in repeated sampling from the conditional set will contain the unknown total Y.

Holt and Smith (1979) provide compelling arguments in favour of conditional design based inference, even though the discussion was confined to one-way post-stratification of a simple random sample in which case it is natural to make inferences conditional on the realized strata sample sizes. Rao (1992, 1994) and Casady and Valliant (1993) studied conditional inference when only the auxiliary total X is known from external sources. In the latter case, conditioning on the NHT estimator $\hat{X}_{NHT}$ may be reasonable because it is "approximately" an ancillary statistic when X is known and the difference $\hat{X}_{NHT} - X$ provides a measure of imbalance in the realized sample. Conditioning on $\hat{X}_{NHT}$ leads to the "optimal" linear regression estimator which has the same form as the GREG estimator (5) with $\hat{B}$ given by (6) replaced by the estimated optimal value $\hat{B}_{opt}$ of the regression coefficient which involves the estimated covariance of $\hat{Y}_{NHT}$ and $\hat{X}_{NHT}$ and the estimated variance of $\hat{X}_{NHT}$. This optimal estimator leads to conditionally valid design-based inferences and model-unbiased under the working model (4). It is also a calibration estimator depending only on the total X and it can be expressed as $\sum_{i\in s} \tilde{w}_i(s) y_i$ with

weights $\tilde{w}_i(s) = d_i \tilde{g}_i(s)$ and the calibration factor $\tilde{g}_i(s)$ depending only on the total X and the sample x-values. It works well for stratified random sampling (commonly used in establishment surveys). However, $\hat{B}_{opt}$ can become unstable in the case of stratified multistage sampling unless the number of sample clusters minus the number of strata is fairly large. GREG estimator does not require the latter condition but it can perform poorly in terms of conditional bias ratio and conditional coverage rates, as shown by Rao (1996). The unbiased NHT estimator can be very bad conditionally unless the design ensures that measure of imbalance as defined above is small. For example, in the Hansen et al. (1983) design based on efficient x-stratification, the imbalance is small and the NHT estimator indeed performed well conditionally.

Tille (1998) proposed an NHT estimator of the total Y based on approximate conditional inclusion probabilities given $\hat{X}_{NHT}$. His method also leads to conditionally valid inferences, but the estimator is not calibrated to X unlike the "optimal" linear regression estimator. Fuller (2002) suggested a calibrated GREG version.

I believe practitioners should pay more attention to conditional aspects of design-based inference and seriously consider the new methods that have been proposed.

Kalton (2002) has given compelling arguments for favouring design-based approaches (possibly model-assisted and/or conditional) for inference on finite population descriptive parameters. Smith (1994) named design-based inference as "procedural inference" and argued that procedural inference is the correct approach for surveys in the public domain.

3. CALIBRATION ESTIMATORS

Calibration weights $w_i(s)$ that ensure consistency with user-specified auxiliary totals X are obtained by adjusting the design weights $d_i = \pi_i^{-1}$ to satisfy the benchmark constraints $\sum_{i \in s} w_i(s) x_i = X$. Estimators that use calibration weights are called calibration estimators and they use a single set of weights $\{w_i(s)\}$ for all the variables of interest. We have noted in section 2.4 that the model-assisted GREG estimator is a calibration estimator, but a calibration estimator may not be model-assisted in the sense it could be model-biased under a working model (4) unless the $x-$variables in the model exactly match the variables corresponding to the user-specified totals. For example, suppose the working model suggested by the data is a quadratic in a scalar variable x while the user-specified total is only its total X. The resulting calibration estimator can perform poorly even in fairly large samples, as noted in section 2.3, unlike the model-assisted GREG estimator based on the working quadratic model that requires the population total of the quadratic variables x_i^2 in addition to X.

Post-stratification has been extensively used in practice to ensure consistency with known cell counts corresponding to a post-stratification variable, for example counts in different age groups ascertained from external sources such as demographic projections. The resulting post-stratified estimator is a calibration estimator. Calibration estimators that ensure consistency with known marginal counts of two or more post-stratification variables have also been employed in practice; in particular raking ratio estimators that

are obtained by benchmarking to the marginal counts in turn until convergence is approximately achieved, typically in four or less iterations. Raking ratio weights $w_i(s)$ are always positive. In the past, Statistics Canada used raking ratio estimators in the Canadian Census to ensure consistency of 2B-item estimators with known 2A-item counts. In the context of the Canadian Census, Brackstone and Rao (1979) studied the efficiency of raking ratio estimators and also derived Taylor linearization variance estimators when the number of iterations is four or less. Raking ratio estimators have also been employed in the U.S. Current Population Survey (CPS). It may be noted that the method of adjusting cell counts to given marginal counts in a two-way table was originally proposed in the landmark paper by Deming and Stephan (1940).

Unified approaches to calibration, based on minimizing a suitable distance measure between calibration weights and design weights subject to benchmark constraints, have attracted the attention of users due to their ability to accommodate arbitrary number of user-specified benchmark constraints, for example, calibration to the marginal counts of several post-stratification variables. Calibration software is also readily available, including GES (Statistics Canada), LIN WEIGHT (Statistics Netherlands), CALMAR (INSEE, France) and CLAN97 (Statistics Sweden).

A chi-squared distance, $\sum_{i\in s} q_i (d_i - w_i)^2 / d_i$, leads to the GREG estimator (5), where the x-vector corresponds to the user-specified benchmark constraints (BC) and $w_i(s)$ is denoted as w_i for simplicity (Huang and Fuller, 1978; Deville and Sarndal, 1992). However, the resulting calibration weights may not satisfy desirable range restrictions (RR), for example some weights may be negative or too large especially when the number of constraints is large and the variability of the design weights is large. Hwang and Fuller (1978) proposed a scaled modified chi-squared distance measure and obtained the calibration weights through an iterative solution that satisfies BC at each iteration. However, a solution that satisfies BC and RR may not exist. Another method, called shrinkage minimization (Singh and Mohl, 1996) has the same difficulty. Quadratic programming methods that minimize the chi-squared distance subject to both BC and RR have also been proposed (Hussain, 1969) but the feasible set of solutions satisfying both BC and RR can be empty. Alternative methods propose to change the distance function (Deville and Sarndal, 1992) or drop some of the BC (Bankier *et al.*, 1992). For example, an information distance of the form $\sum_{i\in s} q_i \{w_i \log(w_i / d_i) - w_i + d_i\}$ gives raking ratio estimators with non-negative weights w_i, but some of the weights can be excessively large. “Ridge” weights obtained by minimizing a penalized chi-squared distance have

also been proposed (Chambers, 1996), but no guarantee that either BC or RR are satisfied, although the weights are more stable than the GREG weights. Rao and Singh (1997) proposed a "ridge shrinkage" iterative method that ensures convergence for a specified number of iterations by using a built-in tolerance specification to relax some BC while satisfying RR. Chen et al. (2002) proposed a similar method.

GREG calibration weights have been used in the Canadian Labour Force Survey and more recently it has been extended to accommodate composite estimators that make use of sample information in previous months, as noted in Section 1 (Fuller and Rao, 2001; Gambino et al., 2001; Singh and Wu, 2001). GREG-type calibration estimators have also been used for the integration of two or more independent surveys from the same population. Such estimators ensure consistency between the surveys, in the sense that the estimators from the two surveys for common variables are identical, as well as benchmarking to known population totals (Renssen and Nieuwenbroek, 1997; Singh and Wu, 1996; Merkouris, 2004). For the 2001 Canadian Census, Bankier (2003) studied calibration weights corresponding to the "optimal" linear regression estimator (section 2.3) under stratified random sampling. He showed that "optimal" calibration method performed better than the GREG calibration, used in the previous census, in the sense of allowing more BC to be retained while at the same time allowing the calibration weights to be at least one. The "optimal" calibration weights can be obtained from GES software by including the known strata sizes in the BC and defining the tuning constant q_i suitably. Note that the "optimal" calibration estimator also has desirable conditional design properties (section 2.4). Weighting for the 2001 Canadian census switched from projection GREG (used in the 1996 census) to "optimal" linear regression.

Demnati and Rao (2004) derived Taylor linearization variance estimators for a general class of calibration estimators with weights $w_i = d_i F(x_i'\hat{\lambda})$, where the LaGrange multiplier $\hat{\lambda}$ is determined by solving the calibration constraints. The choice $F(a) = 1 + a$ gives GREG weights and $F(a) = e^a$ leads to raking ratio weights. In the special case of GREG weights, the variance estimator reduces to $v(ge)$ given in section 2.3.

We refer the reader to the Waksberg award paper of Fuller (Fuller, 2002) for an excellent overview and appraisal of regression estimation in survey sampling, including calibration estimation.

4. UNEQUAL PROBABILITY SAMPLING

We have noted in Section 1 that PPS sampling of PSUs within strata in large-scale surveys was practically motivated by the desire to achieve approximately equal workloads. PPS sampling also achieves significant variance reduction by controlling on the variability arising from unequal PSU sizes without actually stratifying by size. PSUs are typically sampled without replacement such that the PSU inclusion probability, π_i, is

proportional to PSU size measure x_i. For example, systematic PPS sampling, with or without initial randomization of the PSU labels, is an inclusion probability proportional to size (IPPS) design (also called πPS design) that has been used in many complex surveys, including the Canadian LFS. The estimator of a total associated with an IPPS design is the NHT estimator.

Development of suitable (IPPS, NHT) strategies raises theoretically challenging problems, including the evaluation of exact joint inclusion probabilities, π_{ij}, or accurate approximations to π_{ij} requiring only the individual π_is, that are needed in getting unbiased or nearly unbiased variance estimator. My own 1961 Ph.D. thesis at Iowa State University addressed the latter problem. Several solutions, requiring sophisticated theoretical tools, have been published since then by talented mathematical statisticians. However, this theoretical work is often classified as "theory without application" because it is customary practice to treat the PSUs as if sampled with replacement that leads to great simplification. The variance estimator is simply obtained from the estimated PSU totals and in fact this assumption is the basis for re-sampling methods (Section 5). This variance estimator can lead to substantial over-estimation unless the overall PSU sampling fraction is small. The latter may be true in many large-scale surveys, In the following paragraphs, I will try to demonstrate that the theoretical work on (IPPS, NHT) strategies as well as some non-IPPS designs indeed have wide practical applicability.

First, I will focus on (IPPS, NHT) strategies. In Sweden and some other countries in Europe, stratified single-stage sampling is often used because of the availability of list frames and IPPS designs are attract options, but sampling fractions are often large. For example, Rosen (1991) notes that Statistics Sweden's Labour Force Barometer surveys some 100 different populations using systematic PPS sampling and that the sampling rates can be as high as 50% or even more. Aires and Rosen (2005) studied Pareto πPS sampling for Swedish surveys. This method has attractive properties, including fixed sample size, simple sample selection, good estimation precision, consistent variance estimation regardless of sampling rates. It also allows sample coordination through permanent random numbers as in Poisson sampling, but the latter method leads to variable sample size. Because of these merits, Pareto πPS has been implemented in a number of Statistics Sweden surveys, notably in price index surveys. The method of Rao-Sampford (see Brewer and Hanif, 1982, p.28) leads to exact IPPS designs and non-negative unbiased variance estimators for arbitrary fixed sample sizes. It has been implemented in the new version of SAS. Stehman and Overton (1994) note that variable probability structure arises naturally in environmental surveys rather than being selected just for enhanced efficiency, and that the $\pi_i s$ are only known for the units i in the sample s. By treating the sample design as randomized systematic PPS, Stehman and Overton obtained approximations to the π_{ij} s that depend only $\pi_i, i \in s$, unlike the original approximations of Hartley and Rao (1962) that require the sum of squares of all the π_is in the population. In their applications, the sampling rates are substantial to warrant the evaluation of the joint inclusion probabilities.

I will now turn to non-IPPS designs using estimators different from the NHT estimator that ensure zero variance when y is exactly proportional to x. The random group method of Rao, Hartley and Cochran (1962) permits a simple non-negative variance estimator for any fixed sample size and yet compares favourably to (IPPS, NHT) strategies in terms of

efficiency and always more efficient than PPS with replacement strategy. Schabenberger and Gregoire (1994) noted that (IPPS, NHT) strategies have not enjoyed much application in forestry because of difficulty in implementation and recommended the Rao-Hartley-Cochran strategy in view of its remarkable simplicity and good efficiency properties. It is interesting to note that this strategy has been used in the Canadian LFS on the basis of its suitability for switching to new size measures, using the Keyfitz method within each random group. On the other hand, (IPPS, NHT) strategies are not readily suitable for this purpose. I understand that the Rao-Hartley-Cochran strategy is often used in audit sampling and other accounting applications.

Murthy (1957) used a non-IPPS design based on drawing successive units with probabilities $p_i, p_j/(1-p_i), p_k/(1-p_i-p_j)$ and so on, and the following estimator:

$$\hat{Y}_M = \sum_{i \in s} y_i \frac{p(s \mid i)}{p(s)}, \qquad (9)$$

where $p(s \mid i)$ is the conditional probability of obtaining the sample s given that unit i was selected first. He also provided a non-negative variance estimator requiring the conditional probabilities, $p(s \mid i, j)$, of obtaining s given i and j are selected in the first two draws. This method did not receive practical attention for several years due to computational complexity, but more recently it has been applied in unexpected areas, including oil discovery (Andreatta and Kaufman, 1986) and sequential sampling including inverse sampling and some adaptive sampling schemes (Salehi and Seber, 1997). It may be noted that adaptive sampling has received a lot of attention in recent years because of its potential as an efficient sampling method for estimating totals or means of rare populations (Thompson and Seber, 1996). In the oil discovery application, successive sampling scheme is a characterization of discovery and the order in which fields are discovered is governed by sampling proportional to field size and without replacement, following the industry folklore “on the average, the big fields are found first”. Here $p_i = y_i / Y$ and the total oil reserve Y is assumed to be known from geological considerations. In this application, geologists are interested in the size distribution of all fields in the basin and when a basin is partially explored the sample is composed of magnitudes y_i of discovered deposits. The size distribution function $F(a)$ can be estimated by using Murthy’s estimator (9) with y_i replaced by the indicator variable $I(y_i \le a)$. The computation of $p(s \mid i)$ and $p(s)$, however, is formidable even for moderate sample sizes. To overcome this computational difficulty, Andreatta and Kaufman (1986) used integral representations of these quantities to develop asymptotic expansions of Murthy’s estimator, the first few terms of which are easily computable.

Similarly, they obtain computable approximations to Murthy's variance estimator. Note that the NHT estimator of $F(a)$ is not feasible here because the inclusion probabilities are functions of all the y – values in the population.

The above discussion is intended to demonstrate that a particular theory can have applications in diverse practical areas even if it is not needed in a particular situation, such as large-scale surveys with negligible first stage sampling fractions. Also it shows that unequal probability sampling designs play a vital role in survey sampling, despite Sarndal's (1996) contention that simpler designs, such as stratified SRS and stratified Bernoulli sampling, together with GREG estimators should replace strategies based on unequal probability sampling without replacement.

5. ANALYSIS OF SURVEY DATA AND RESAMPLING METHODS

Standard methods of data analysis are generally based on the assumption of simple random sampling, although some software packages do take account of survey weights and provide correct point estimates. However, application of standard methods to survey data, ignoring the design effect due to clustering and unequal probabilities of selection, can lead to erroneous inferences even for large samples. In particular, standard errors of parameter estimates and associated confidence intervals can be seriously under-stated, type I error rates of tests of hypotheses can be much bigger than the nominal levels, and standard model diagnostics, such as residual analysis to detect model deviations, are also affected. Kish and Frankel (1974) and others drew attention to some of those problems and emphasized the need for new methods that take proper account of the complexity of data derived from large-scale surveys. Fuller (1975) developed asymptotically valid methods for linear regression analysis, based on Taylor linearization variance estimators. Rapid progress has been made over the past 20 years or so in developing suitable methods. Re-sampling methods play a vital role in developing methods that take account of survey design in the analysis of data. All one needs is a data file containing the observed data, the final survey weights and the corresponding final weights for each pseudo-replicate generated by the re-sampling method. Software packages that take account of survey weights in the point estimation of parameters of interest can then be used to calculate the correct estimators and standard errors, as demonstrated below. As a result, re-sampling methods of inference have attracted the attention of users as they can perform the analyses themselves very easily using standard software packages. However, releasing public-use data files with replicate weights can lead to confidentiality issues, such as the identification of clusters from replicate weights. In fact, at present a challenge to theory is to develop suitable methods that can preserve confidentiality of the data. Lu, Brick and Sitter (2004) proposed grouping strata and then form pseudo-replicates using the combined strata for variance estimation, thus limiting the risk of cluster identification from the resulting public-use data file. A method of inverse sampling to undo the complex survey data structure and yet provide protection against revealing cluster labels (Hinkins, Oh and Scheuren, 1997; Rao, Scott and Benhin, 2003) appears promising, but

much work on inverse sampling methods remains to be done before it becomes attractive to the user.

Rao and Scott (1981, 84) made a systematic study of the impact of survey design effect on standard chi-squared and likelihood ratio tests associated with a multi-way table of estimated counts of proportions. They showed that the test statistic is asymptotically distributed as a weighted sum of independent χ_1^2 variables, where the weights are the eigenvalues of a "generalized design effects" matrix. This general result shows that the survey design can have a substantial impact on the type I error rate. Rao and Scott proposed simple first-order corrections to the standard chi-squared statistics that can be computed from published tables that include estimates of design effects for cell estimates and their marginal totals, thus facilitating secondary analyses from published tables. They also derived second order corrections that are more accurate, but require the knowledge of a full estimated covariance matrix of the cell estimates, as in the case of familiar Wald tests. However, Wald tests can become highly unstable as the number of cells in a mult-way table increases and the number of sample clusters decrease, leading to unacceptably high type I error rates compared to the nominal levels, unlike the Rao-Scott second order corrections (Thomas and Rao, 1987). The first and second order corrections are now known as Rao-Scott corrections and are given as default options in the new version of SAS. Roberts, Rao and Kumar (1987) developed Rao-Scott type corrections to tests for logistic regression analysis of estimated cell proportions associated with a binary response variable. They applied the methods to a two-way table of employment rates from the Canadian LFS, 1977 obtained by cross-classifying age and education groups. Bellhouse and Rao (2002) extended the work of Roberts et al. to the analysis of domain means using generalized linear models. They applied the methods to domain means from a Fiji Fertility Survey cross-classified by education and years since the woman's first marriage, where a domain mean is the mean number of children even born for women of Indian race belonging to the domain.

Re-sampling methods in the context of large -scale surveys using stratified multi-stage designs have been studied extensively. For inference purposes, the sample PSUs are treated as if drawn with replacement within strata. This leads to over-estimation of variances but it is small if the overall PSU sampling fraction is negligible. Let $\hat{\theta}$ be the survey-weighted estimator of a "census" parameter of interest computed from the final weights w_i, and let the corresponding weights for each pseudo-replicate r generated by the re-sampling method be denoted by $w_i^{(r)}$. The estimator based on the pseudo-replicate

weights $w_i^{(r)}$ is denoted as $\hat{\theta}^{(r)}$ for each $r = 1,..., R$. Then a re-sampling variance estimator of $\hat{\theta}$ is of the form

$$v(\hat{\theta}) = \sum_{r=1}^{R} c_r (\hat{\theta}^{(r)} - \hat{\theta})(\hat{\theta}^{(r)} - \hat{\theta})' \qquad (10)$$

for specified coefficients c_r in (10) determined by the re-sampling method.

Commonly used re-sampling methods include (a) delete-cluster (delete-PSU) jackknife, (b) balanced repeated replication (BRR) particularly for $n_h = 2$ PSUs in each stratum h and (c) Rao-Wu (1988) bootstrap. Jackknife pseudo-replicates are obtained by deleting each sample cluster $r = (hj)$ in turn, leading to jackknife design weights $d_i^{(r)}$ taking the value 0 if the sample unit i is in the deleted cluster, $n_h d_i /(n_h - 1)$ if i is not in the deleted cluster but in the same stratum, and unchanged if i is in a different stratum. The jackknife design weights are then adjusted for unit non-response and post-stratification, leading to the final jackknife weights $w_i^{(r)}$. The jackknife variance estimator is given by (10) with $c_r = (n_h - 1)/n_h$ when $r = (hj)$. The delete-cluster jackknife method has two possible disadvantages: (1) When the total number of sampled PSUs, $n = \sum n_h$, is very large, R is also very large because $R = n$. (2) It is not known if the delete-jackknife variance estimator is design-consistent in the case of non-smooth estimators $\hat{\theta}$, for example the survey-weighted estimator of the median. For simple random sampling, the jackknife is known to be inconsistent for the median or other quantiles. It would be theoretically challenging and practically relevant to find conditions for the consistency of delete-cluster jackknife variance estimator of a non-smooth estimator $\hat{\theta}$.

BRR can handle non-smooth $\hat{\theta}$, but it is readily applicable only for the important special case of $n_h = 2$ PSUs per stratum. A minimal set of balanced half-samples can be constructed from an $R \times R$ Hadamard matrix by selecting H columns, excluding the column of $+1$'s, where $H + 1 \le R \le H + 4$ (McCarthy, 1969). The BRR design weights $d_i^{(r)}$ equal $2d_i$ or 0 according as whether or not i is in the half-sample. A modified BRR, due to Bob Fay, uses all the sampled units in each replicate unlike the BRR by defining the replicate design weights as $d_i^{(r)}(\varepsilon) = (1+\varepsilon)d_i$ or $(1-\varepsilon)d_i$ according as whether or not i is in the half-sample, where $0 < \varepsilon < 1$; a good choice of ε is $\frac{1}{2}$. The modified BRR weights are then adjusted for non-response and post-stratification to get the final weights $w_i^{(r)}(\varepsilon)$ and the estimator $\hat{\theta}^{(r)}(\varepsilon)$. The modified BRR variance estimator is given by (10) divided by ε^2 and $\hat{\theta}^{(r)}$ replaced by $\hat{\theta}^{(r)}(\varepsilon)$, see Rao and Shao (1999). The modified BRR is particularly useful under independent re-imputation for missing item responses in each replicate because it can use the donors in the full sample to impute unlike the BRR that uses the donors only in the half-sample.

The Rao-Wu bootstrap is valid for arbitrary $n_h (\geq 2)$ unlike the BRR, and it can also handle non-smooth $\hat{\theta}$. Each bootstrap replicate is constructed by drawing a simple random sample of PSUs of size $n_h - 1$ from the n_h sample clusters, independently across the strata. The bootstrap design weights $d_i^{(r)}$ are given by $[n_h/(n_h - 1)]m_{hi}^{(r)} d_i$ if i is in

stratum h and replicate r, where $m_{hi}^{(r)}$ is the number of times sampled PSU (hi) is selected, $\sum_i m_{hi}^{(r)} = n_h - 1$. The weights $d_i^{(r)}$ are then adjusted for unit non-response and – post-stratification to get the final bootstrap weights and the estimator $\hat{\theta}^{(r)}$. Typically, $R = 500$ bootstrap replicates are used in the bootstrap variance estimator (10). Several recent surveys at Statistics Canada have adopted the bootstrap method for variance estimation because of the flexibility in the choice of R and wider applicability. Users of Statistics Canada micro survey data files seem to be very happy with the bootstrap method for analysis of data.

Early work on the jackknife and the BRR was largely empirical (e.g., Kish and Frankel, 1974). Krewski and Rao (1981) formulated a formal asymptotic framework appropriate for stratified multi-stage sampling and established design consistency of the jackknife and BRR variance estimators when $\hat{\theta}$ can be expressed as a smooth function of estimated means. Several extensions of this basic work have been reported in the recent literature; see the book by Shao and Tu (1995, Chapter 6). Theoretical support for re-sampling methods is essential for their use in practice.

In the above discussion, I simply denoted $\hat{\theta}$ as the estimator of a "census" parameter. Typically, the census parameter θ_C is motivated by an underlying super-population model and the census is regarded as a sample generated by the model, leading to census estimating equations whose solution is θ_C. The census estimating functions $U_C(\theta)$ are simply population totals of functions $u_i(\theta)$ with zero expectation under the assumed model, and the census estimating equations are given by $U_C(\theta) = 0$. Kish and Frankel (1974) argued that the census parameter makes sense even if the model is not correctly specified. For example, in the case of linear regression, the census regression coefficient could explain how much of the relationship between the response variable and the independent variables is accounted by a linear regression model. Noting that the census estimating functions are simply population totals, survey weighted estimators $\hat{U}(\theta)$ from the full sample and $\hat{U}^{(r)}(\theta)$ from each pseudo-replicate are obtained. The solutions of corresponding estimating equations $\hat{U}(\theta) = 0$ and $\hat{U}^{(r)}(\theta) = 0$ give $\hat{\theta}$ and $\hat{\theta}^{(r)}$ respectively. Note that the re-sampling variance estimators are designed to estimate the variance of $\hat{\theta}$ as an estimator of the census parameters but not the model parameters. Under certain conditions, the difference can be ignored but in general we have a two-phase sampling situation, where the census is the first phase sample from the super-

population and the sample is a probability sample from the census population. Recently, some useful work has been done on two-phase variance estimation when the model parameters are the target parameters (Graubard and Korn, 2002; Rubin-Bleuer and Schiopu Kratina, 2004), but more work is needed to address the difficulty in specifying the covariance structure of the model errors.

A difficulty with the bootstrap is that the solution $\hat{\theta}^{(r)}$ may not exist for some bootstrap replicates r (Binder, Kovacevic and Roberts, 2004). Rao and Tausi (2004) used estimating function (EF) bootstrap method that avoids the difficulty. In this method, we solve $\hat{U}(\theta) = \hat{U}^{(r)}(\hat{\theta})$ for θ using only one-step of the Newton-Raphson iteration with $\hat{\theta}$ as the starting value. The resulting estimator $\tilde{\theta}^{(r)}$ is then used in (10) to get the EF bootstrap variance estimator of $\hat{\theta}$ which can be readily implemented from the data file providing replicate weights, using slight modifications of any software package that accounts for survey weights. It is interesting to note that the EF bootstrap variance estimator is equivalent to a Taylor linearization sandwich variance estimator that uses the bootstrap variance estimator of $\hat{U}(\theta)$ and the inverse of the observed information matrix (derivative of $-\hat{U}(\theta)$), both evaluated at $\theta = \hat{\theta}$ (Binder, Kovacevic and Roberts, 2004).

Pfeffermann (1993) discussed the role of design weights in the analysis of survey data. If the population model holds for the sample (i.e., if there is no sample selection bias), then model-based unweighted estimators will be more efficient than the weighted estimators and lead to valid inferences, especially for data with smaller sample sizes and larger variation in the weights. However, for typical data from large-scale surveys, the survey design is informative and the population model may not hold for the sample. As a result, the model-based estimators can be seriously biased and inferences can be erroneous. Pfeffermann and his colleagues initiated a new approach to inference under informative sampling; see Pfeffermann and Sverchkov (2003) for recent developments. This approach seems to provide more efficient inferences compared to the survey weighted approach, and it certainly deserves the attention of users of survey data. However, much work remains to be done, especially in handling data based on multi-stage sampling.

6. SMALL AREA ESTIMATON

Previous sections of this paper have focussed on traditional methods that use direct domain estimators based on domain-specific sample observations along with auxiliary population information. Such methods, however, may not provide reliable inferences when the domain sample sizes are very small or even zero for some domains. Domains or sub-populations with small or zero sample sizes are called small areas in the literature.

Demand for reliable small area statistics has greatly increased in recent years because of the growing use of small area statistics in formulating policies and programs, allocation of funds and regional planing. Clearly, it is seldom possible to have a large enough overall sample size to support reliable direct estimates for all domains of interest. Also, in practice, it is not possible to anticipate all uses of survey data and "the client will always require more than is specified at the design stage (Fuller, 1999, p.344). In making

estimates for small areas with adequate level of precision, it is often necessary to use"indirect" estimators that borrow information from related domains through auxiliary information, such as census and current administrative data, to increase the "effective" sample size within the small areas.

It is now generally recognized that explicit models linking the small areas through auxiliary information and accounting for residual between area variation through random small area effects are needed in developing indirect estimators. Success of such model-based methods heavily depends on the availability of good auxiliary information and thorough validation of models through internal and external evaluations. Many of the methods used in mainstream statistical theory involving models with random effects are relevant to small area estimation, including empirical best (or Bayes), empirical best linear unbiased prediction and hierarchical Bayes based on prior distributions on the model parameters. A comprehensive account of such methods is given in my 2003 Wiley book (Rao, 2003). Practical relevance and theoretical interest of small area estimation have attracted the attention of many researchers, leading to important advances in point estimation as well as measuring the variability of the estimators. The "new" methods have been applied successfully worldwide to a variety of small area problems. . Model-based methods have been recently used to produce current county and school district estimates of poor school-age children in U.S.A. Based on the model-based estimates, the U.S. Department of Education allocates annually over $7 billion of funds to counties. The allocated funds support compensatory education programs to meet the needs of educationally disadvantaged children. We refer to Rao (2003, example 7.1.2) for details of this application. In the United Kingdom the Office of National Statistics established a Small Area Estimation Project to develop model-based estimates at the level of political wards (roughly 2000 households). The practice and estimation methods of U.S. federal statistical programs that use indirect estimators to produce published estimates are documented in Schaible (1996).

Small area estimation is a striking example of the interplay between theory and practice. Theoretical advances are impressive, but many practical issues need further attention of theory. Such issues include: (a) Benchmarking model-based estimators to agree with reliable direct estimators at large area levels. (b) Developing and validating suitable linking models and addressing issues such as errors in variables, incorrect specification of the linking model and omitted variables. (c) Development of methods that satisfy multiple goals: good area-specific estimates, good ranks and a good histogram of small areas.

7. SOME THEORY DESERVING ATTENTION OF PRACTICE

In this section, I will briefly mention some examples of important theory that exists but not widely used in practice.

7.1. Empirical Likelihood Inference

Traditional sampling theory largely focused on point estimation and associated standard errors, appealing to normal approximations for confidence intervals on parameters of interest. In the mainstream statistics, the empirical likelihood (EL) approach (Owen 1988) has attracted a lot of attention due to several desirable properties. It provides a non-parametric likelihood, leading to EL ratio confidence intervals similar to the parametric likelihood ratio intervals. The shape and orientation of EL intervals are determined entirely by the data, and the intervals are range preserving and transformation respecting, and are particularly useful in providing balanced tail error rates, unlike the symmetric normal theory intervals. As noted in Section 1.1, the EL approach was in fact first introduced in the sample survey context by Hartley and Rao (1968), but their focus was on inferential issues related to point estimation. Chen, Chen and Rao (2003) obtained EL intervals on the population mean under simple random and stratified random sampling for populations containing many zeros. Such populations are encountered in audit sampling, where y denotes the amount of money owed to the government and the mean $\bar{Y}$ is the average amount of excessive claims. Previous work on audit sampling used parametric likelihood ratio intervals based on parametric mixture distributions for the variable y. Such intervals perform better than the standard normal theory intervals, but EL intervals perform better under deviations from the assumed mixture model, by providing non-coverage rate below the lower bound closer to the nominal error rate and also larger lower bound. For general designs, Wu and Rao (2004) used a pseudo-empirical likelihood (Chen and Sitter, 1999) to obtain adjusted pseudo-EL intervals on the mean and the distribution function that account for the design features, and showed that the intervals provide more balanced tail error rates than the normal theory intervals. The EL method also provides a systematic approach to calibration estimation and integration of surveys. We refer the reader to the review papers by Rao (2004) and Wu and Rao (2005). Further refinements and extensions remain to be done, particularly on the pseudo-empirical likelihood, but the EL theory in the survey context deserves the attention of practice.

7.2 Exploratory Analyses of Survey Data

In Section 5 we discussed methods for confirmatory analysis of survey data taking design into account, such as point estimation of model (or census) parameters and associated standard errors and formal tests of hypotheses. Graphical displays and exploratory data analyses of survey data are also very useful. Such methods have been extensively

developed in the mainstream literature. Only recently, some extensions of these modern methods are reported in the survey literature and deserve the attention of practice. I will briefly mention some of those developments. First, non-parametric kernel density estimates are commonly used to display the shape of a data set without relying on parametric models. They can also be used to compare different sub-populations.

Bellhouse and Stafford (1999) provided kernel density estimators that take account of the surey design and studied their properties and applied the methods to data from the Ontario Health Survey. Buskirk and Lohr (2005) studied asymptotic and finite sample properties of kernel density estimators and obtained confidence bands. They applied the methods to data from the US National Crime Victimization Survey and the US National Health and Nutrition Examination Survey.

Secondly, Bellhouse and Stafford (2001) developed local polynomial regression methods, taking design into account, that can be used to study the relationship between a response variable and predictor variables, without making strong parametric model assumptions. The resulting graphical displays are useful in understanding the relationships and also for comparing different sub-populations. Bellhouse and Stafford (2001) illustrated local polynomial regression on the Ontario Health Survey data; for example, the relationship between body mass index of females and age. Bellhouse, Chipman and Stafford (200-) studied additive models for survey data via penalized least squares method to handle more than one predictor variable, and illustrated the methods on the Ontario Health Survey data. This approach has many advantages in terms of graphical display, estimation, testing and selection of “smoothing” parameters for fitting the models.

7.3 Measurement Errors

Typically, measurement errors are assumed to be additive with zero means. As a result, usual estimators of total and means remain unbiased or consistent. However, this nice feature may not hold for more complex parameters such as distribution function, quantiles and regression coefficients. In the latter case, usual estimators will be biased, even for large samples, and hence can lead to erroneous inferences (Fuller, 1995). It is possible to obtain bias-adjusted estimators if estimates of measurement error variances are available. The latter may be obtained by allocating resources at the design stage to make repeated observations on a sub-sample. Fuller (1975, 1995) has been a champion of proper methods in the presence of measurement errors and the bias-adjusted methods deserve the attention of practice.

Hartley and Rao (1968) and Hartley and Biemer (1981) provided interviewer and coder assignment conditions that permit the estimation of sampling and response variances for the mean or total from current surveys. Unfortunately, current surveys are often not designed to satisfy those conditions and even if they do the required information on interviewer and coder assignements is seldom available at the estimation stage.

Linear components of variance models are often used to estimate interviewer variability. Such models are appropriate for continuous response but not for binary responses. The linear model approach for binary responses can result in underestimating the intra-interviewer correlations. Scott and Davis (2001) proposed multi-level models for binary responses to estimate interviewer variability. Given that responses are often binary in many surveys, practice should pay attention to such models for proper analyses of survey data with binary responses.

7.4 Imputation for Missing Survey Data

Imputation is commonly used in practice to fill in missing item values. It ensures that the results obtained from different analyses of the completed data set are consistent with one another by using the same survey weight for all items. Marginal imputation methods, such as ratio, nearest neighbour and random donor within imputation classes are used by many statistical agencies. Unfortunately, the imputed values are often treated as if they were true values and then compute estimates and variance estimates. The imputed point estimates of marginal parameters are generally valid under an assumed response mechanism or imputation model. But the "naïve" variance estimators can lead to erroneous inferences even for large samples; in particular, serious underestimation of the variance of the imputed estimator because the additional variability due to estimating the missing values is not taken into account. Advocates of Rubin's (1987) multiple imputation claim that the multiple imputation variance estimator can fix this problem because a between imputed estimators sum of squares is added to the average of naïve variance estimators resulting from the multiple imputations. Unfortunately, there are difficulties with multiple imputation variance estimators, as discussed by Kott (1995), Fay (1996), Binder and Sun (1996), Wang and Robins (1998), Kim, Brick and Fuller (2004) and others. Moreover, single imputation is often preferred due to operational and cost considerations. Some impressive advances have been made in recent years on making efficient and valid inferences from singly imputed data sets. We refer the reader to review papers by Shao (2002) and Rao (2000, 2005) for methods of variance estimation under single imputation that deserve the attention of practice. But much remains to be done.

8. CONCLUDING REMARKS

Joe Waksberg's contributions to sample survey theory and methods truly reflect the interplay between theory and practice. Working at the US Census Bureau and later at Westat, he faced real practical problems and often produced beautiful theory to solve them. For example, his landmark paper (Waksberg, 1978) produced an ingenious method for random digit dialing that significantly reduced the survey costs compared to dialing

numbers completely at random. He provided sound theory to demonstrate its efficiency. I feel greatly honoured to receive the 2004 Waksberg award for survey methodology.

REFERENCES

AIRES, N. and ROSÉN, B. (2005). On inclusion probabilities and relative estimator bias for Pareto π ps sampling. *Journal of Statistical Planning and Inference*, 128, 543-567.

ANDREATTA, G. and KAUFMANN, G.M. (1986). Estimation of finite population properties when sampling is without replacement and proportional to magnitude. *Journal of the American Statistical Association*, 81, 657-666.

BANKIER, M.D. (2003). 2001 Canadian Census weighting: switch from projection GREG to pseudo-optimal regression estimation. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, Technical Report no. 386, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa.

BANKIER, M.D., RATHWELL, S. and MAJKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Working Paper – Methodology Branch, Census Operations Section, Social Survey Methods Division, Statistics Canada, Ottawa.

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.

BELLHOUSE, D.R. and RAO, J.N.K. (2002). Analysis of domain means in complex surveys. *Journal of Statistical Planning and Inference*, 102, 47-58.

BELLHOUSE, D.R. and STAFFORD, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.

BELLHOUSE, D.R. and STAFFORD, J.E. (2001). Local polynomial regression in complex surveys. *Survey Methodology*, 27, 197-203.

BELLHOUSE, D.R., CHIPMAN, H.A. and STAFFORD, J.E. (2004). Additive models for survey data via penalized least squares. Technical Report.

BINDER, D., KOVACEVIC, M. and ROBERTS, G. (2004). Design-based methods for survey data: Alternative uses of estimating functions. *Proceedings of the section on survey research methods*, American Statistical Association.

BINDER, D.A. and SUN, W. (1996). Frequency valid multiple imputation for surveys with a complex design. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 281-286.

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

BRACKSTONE, G. and RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Series C, 42, 97-114.

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.

BUSKIRK, T.D. and LOHR, S.L. (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference*, 128, 165-190.

CASADY, R.J. and VALLIANT, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodology*, 19, 183-192.

CHAMBERS, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

CHEN, J. and SITTER, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 12, 1223-1239.

CHEN, J., CHEN, S.Y. and RAO, J.N.K. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics*, 31, 53-68.

CHEN, J., SITTER, R.R. and WU, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.

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. (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. *Journal of Agricultural Science*, 30, 262-275.

COCHRAN, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 191-212.COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistica*, 17, 164-177.

COCHRAN, W.G. (1953). *Sampling Techniques*. New York: Wiley.

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

DEMING, W.E. and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected margins are known. *The Annals of Mathematical Statistics*, 11, 427-444.

DEMNATI, A. and RAO, J.N.K. (2004). Linearization variance estimators for survey data. *Survey ethodology*, 30, 17-26.

DEVILLE, J. and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DURBIN, J. (1968). Sampling theory for estimates based on fewer individuals than the number selected. *Bulletin of the International Statistical Institute*, 36, No. 3, 113-119.

FAY, R.E. (1996). Alternative peradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

FRANCISCO, C.A. and FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

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

FULLER, W.A. (1995). Estimation in the presence of measurement error. *International Statistical Review*, 63, 121-147.

FULLER, W.A. (1999). Environmental surveys over time, *Journal of Agricultural, Biological and Environmental Statistics*, 4, 331-345.

FULLER, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.

FULLER, W.A. and RAO, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 45-51.

GAMBINO, J., KENNEDY, B. and SINGH, M.P. (2001). Regression composite estimation for the Canadian labour force survey: Evaluation and implementation. *Survey Methodology*, 27, 65-74.

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.

GRAUBARD, B.I. and KORN, E.L. (2002). Inference for superpopulation parameters under sample surveys. *Statistical Science*, 17, 73-96.

HACKING, I. (1975). *History and Philosophy of Science Seminar*.

HAJÉK, J. (1971). Comments on a paper by Basu, D. In *Foundations of Statistical Inference* (Eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston,

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). *Sample Survey Methods and Theory*, Vols. I and II. New York: Wiley.

HANSEN, M.H., HURWITZ, W.N. 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., HURWITZ, W.N., MARKS, E.S. and MAULDIN, W.P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190.

HANSEN, M.H., HURWITZ, W.N., NISSESLSON, H. and STEINBERG, J. (1955). The redesign of the census current population survey. *Journal of the American Statistical Association*, 50, 701-719.

HARTLEY, H.O. (1959). Analytical studies of survey data. In Volume in Honour of Corrado Gini, Instituto di Statistica, Rome, 1-32.

HARTLEY, H.O. and BIEMER, P. (1978). The estimation of nonsampling variance in current surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 257-262.

HARTLEY, H.O. and RAO, J.N.K. (1962). Sampling with unequal probability and without replacement. *The Annals of Mathematical Statistics*, 33, 350-374.

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

HARTLEY, H.O. and RAO, J.N.K. (1978). The estimation of nonsampling variance components in sample surveys. In *Survey Measurement* (Ed. N.K. Namboodiri), New York: Academic Press, 35-43.

HINKINS, S., OH, H.L. and SCHEUREN, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 11-21.

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

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.

HUANG, E.T. and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the social statistics section*, American Statistical Association, 300-305.

HUBBACK, J.A. (1927). Sampling for rice yield in Bihar and Orissa. Imperial Agricultural Research Institute, Pusa, Bulletin No. 166 (represented in *Sankhyā*, 1946, vol. 7, 281-294).

HUSSAIN, M. (1969). Construction of regression weights for estimation in sample surveys. Unpublished M.S. thesis, Iowa State University, Ames, Iowa.

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

KALTON, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics, 18, 129-154.*

KEYFITZ, N. (1951). Sampling with probabilities proportional to size: adjustment for changing in the probabilities. *Journal of the American Statistical Association*, 46, 105-109.

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

KIM, J.K., BRICK, J.M.; FULLER, W.A. and KALTON, G. (2004). On the bias of the multiple imputation variance estimator in survey sampling. Technical Report.

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

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

KOTT, P.S. (1995). A paradox of multiple imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 384-389.

KOTT, P.S. (2005). Randomized-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 129, 263-277.

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

LAPLACE, P.S. (1820). A philosophical essay on probabilities. English translation, Dover, 1951.

LU, W.W., BRICK, M. and SITTER, R.R. (2004). Algorithms for constructing combined strata grouped jackknife and balanced repeated replication with domains. Technical Report, Westat, Rockville, Maryland.

MAHALANOBIS, P.C. (1944). On large scale sample surveys. *Philosophical Transactions of the Royal Society*, London, Series B, 231, 329-451.

MAHALANOBIS, P.C. (1946a). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

MAHALANOBIS, P.C. (1946b). Sample surveys of crop yields in India. *Sankhyā*, 7 269-280.

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

MERKOURIS, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

MURTHY, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18, 379-390.

MURTHY, M.N. (1964). On Mahalanobis' contributions to the development of sample survey theory and methods. In *Contributions to Statistics*: Presented to Professor P.C. Mahalanobis on the occasion of his 70th birthday, Calcutta, Statistical Publishing Society: pp. 283-316.

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

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.

OWEN, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

OWEN, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

OWEN, A.B. (2001). Empirical Likelihood. New York: Chapman and Hall.

OWEN, A.B. (2002). *Empirical Likelihood*. New York: Chapman & Hall/CRC.

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

PFEFFERMANN, D. and SVERCHKOV, M. (2003). Fitting generalized linear models under informative sampling. In *Analysis of Survey Data* (Eds. R.L. Chambers and C.J. Skinner), Chichester: Wiley.

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

RAO, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā*, series A, 28, 47-60.

RAO, J.N.K. (1992). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Proceedings of the workshop on uses of auxiliary information in surveys*, Statistics Sweden.

RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

RAO, J.N.K. (1996). Developments in sample survey theory. *The Canadian Journal of Statistics*, 25, 1-21.

RAO, J.N.K. (1996). Developments in sample survey theory: an appraisal. *The Canadian Journal of Statistics*, 25, 1-21.

RAO, J.N.K. (2000). Variance estimation in the presence of imputation for missing data. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 599-608.

RAO, J.N.K. (2003). *Small Area Estimation*. Hoboken: Wiley.

RAO, J.N.K. (2005). Re-sampling variance estimation with imputed survey data: overview. *Bulletin of the International Statistical Institute*.

RAO, J.N.K. and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

RAO, J.N.K. and SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

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

RAO, J.N.K. and SHAO, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

RAO, J.N.K. and SINGH, A.C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the section on survey research methods*, American Statistical Association, 57-64.

RAO, J.N.K. and TAUSI, M. (2004). Estimating function jackknife variance estimators under stratified multistage sampling. *Communications in Statistics – Theory and Methods*, 33, 2087-2095.

RAO, J.N.K. and WU, C.F.J. (1987). Methods for standard errors and confidence intervals from sample survey data: some recent work. *Bulletin of the International Statistical Institute*,

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., HARTLEY, H.O. and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, series B, 24, 482-491.

RAO, J.N.K., JOCELYN, W. and HIDIROGLOU, M.A. (2003). Confidence interval coverage properties for regression estimators in uni-phase and two-phase sampling. *Journal of Official Statistics*, 19,

RAO, J.N.K., SCOTT, A.J. and BENHIN, E. (2003). Undoing complex survey data structures: Some theory and applications of inverse sampling. *Survey Methodology*, 29, 107-128.

RENSSEN, R.H. and NIEUWENBROEK, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-375.

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

ROSÉN, B. (1991). Variance estimation for systematic pps-sampling.

ROYAL, 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. (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 CUMBERLAND, W.G. (1981). An empirical study of the ratio estimate and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.

RUBIN, D.B. (1987). *Multiple imputation for nonresponse in surveys. New York: Wiley*.

RUBIN-BLEUER, S. and SCHIOPU KRATINA, I. (2004). On the two-phase framework for joint model and design-based inference. Technical Report, Statistics Canada, Ottawa.

SARNDAL, C.-E. (1996). Efficient estimators with variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.

SARNDAL, C.-E., SWENSON, B. and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

SARNDAL, C.-E., SWENSON, B. and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SALEHI, M. and SEBER, G.A.F. (1997). Adaptive cluster sampling with networks selected without replacements, *Biometrika*, 84, 209-219.

SCHABENBERGER, O. and GREGOIRE, T.G. (1994). Competitors to genuine π ps sampling designs. *Survey Methodology*, 20, 185-192.

SCHAIBLE, W.L. (Ed.) (1996). Indirect Estimation in U.S. Federal Programs. New York: Springer

SCOTT, A. and DAVIS, P. (2001). Estimating interviewer effects for survey responses. Proceedings of Statistics Canada Symposium 2001.

SHAO, J. (2002). Resampling methods for variance estimation in complex surveys with a complex design. In *Survey Nonresponse* (Eds. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), New York: Wiley, 303-314.

SHAO, J. and TU, D. (1995). *The Jackknife and the Bootstrap*. New York: Springer Verlag.

SINGH, A.C. and MOHL, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.

SINGH, A.C. and WU, S. (1996). Estimation for multiframe complex surveys by modified regression. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 69-77.

SITTER, R.R. and WU, C. (2001). A note on Woodruff confidence interval for quantiles. *Statistics & Probability Letters*, 55, 353-358.

SMITH, T.M.F. (1994). Sample surveys 1975-1990; an age of reconciliation? *International Statistical Review*, 62, 5-34.

STEHMAN, S.V. and OVERTON, W.S. (1994). Comparison of variance estimators of the Horvitz Thompson estimator for randomized variable probability systematic sampling. *Journal of the American Statistical Association*, 89, 30-43.

SUKHATME, P.V. (1947). The problem of plot size in large-scale yield surveys. *Journal of the American Statistical Association*, 42, 297-310.

SUKHATME, P.V. (1954). *Sampling Theory of Surveys, with Applications.* Ames: Iowa State College Press.

SUKHATME, P.V. and PANSE, V.G. (1951). Crop surveys in India – II. *Journal of the Indian Society of Agricultural Statistics*, 3, 97-168.

SUKHATME, P.V. and SETH, G.R. (1952). Non-sampling errors in surveys. *Journal of the Indian Society of Agricultural Statistics*, 4, 5-41.

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.

THOMPSON, S.K. and SEBER, G.A.F. (1996). *Adaptive Sampling*. New York: Wiley.

TILLÉ, Y. (1998). Estimation in surveys using conditional inclusion probabilities: simple random sampling. *International Statistical Review*, 66, 303-322.

VALIANT, R., DORFMAN, A.H. and ROYALL, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.

WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

WANG, N. and ROBINS, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.

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

WU, C. and RAO, J.N.K. (2004). Empirical likelihood ratio confidence intervals for complex surveys. Submitted for publication.

WU, C. and RAO, J.N.K. (2005). Empirical likelihood approach to calibration using survey data. Paper presented at the 2005 International Statistical Institute meetings, Sydney, Australia.

YATES, F. (1949). *Sampling Methods for Censuses and Surveys*. London: Griffin.

ZARKOVIC, S.S. (1956). Note on the history of sampling methods in Russia. *Journal of the Royal Statistical Society*, Series A, 119, 336-338.