Sample Survey Theory and Methods: Past, Present and Future Directions

J. N. K. Rao*
Carleton University, Ottawa, Canada – jrao34@rogers.com

Wayne A. Fuller*
Iowa State University, Ames, U. S. A. – waf@iastate.edu

Abstract

We discuss important developments in sample survey theory and methods covering the past 100 years. Neyman's 1934 landmark paper laid the theoretical foundations for the probability sampling approach to inference from survey samples. Classical sampling books by Cochran, Hansen, Hurwitz and Madow, Sukhatme and Yates, which appeared in the early 1950s, expanded and elaborated the theory of probability sampling, emphasizing unbiasedness, model free features, and optimal designs that minimize variance for a fixed cost. During the period 1960-1970, theoretical foundations of inference from survey data received attention, with the model-dependent approach generating considerable discussion. Introduction of general purpose statistical software led to the use of such software with survey data, which led to the design of methods specifically for complex survey data. At the same time, weighting methods, such as regression estimation and calibration, became practical and design consistency replaced unbiasedness as the requirement for standard estimators. A bit later computer-intensive resampling methods also became practical for large scale survey samples. Improved computer power led to more sophisticated imputation for missing data, use of more auxiliary data, and more complex estimation procedures in general. Longitudinal surveys became more common and some treatment of measurement errors beyond the effects to reduce the effects occurred. The most notable use of models was in the expanded use of small area estimation. Future directions in research and methods will be influenced by budgets, response rates, timeliness, improved data collection devices, and availability of auxiliary data, some of which will be called "big data". Most importantly, survey taking will be impacted by changing cultural behavior and by a changing physical-technical environment.

Keywords: History of survey sampling; data collection; probability sampling; survey inference

1. Introduction

The topic "sample surveys" is too broad to address in a short talk or in a long book. To reduce the topic, we divide samples into two categories: "probability based" and "other". Probability samples are described in classic texts such as Cochran (1977) and are such that every element of the finite population of interest has a known probability of entering the sample. Note that the sample is selected from a specified finite population.

"Other" samples are much more widespread and defy enumeration, but include convenience samples, quota samples, Web surveys, etc. We will emphasize probability based sampling. Our discussion is most relevant for samples of human respondents and for large general purpose samples, the surveys of our greatest experience. Likewise our experience with application is concentrated in Canada and the United States. We do not pretend completeness and recognize that we have omitted important topics of current interest..

2. Early Landmark Contributions: 1920-1960

Kiaer (1897) is perhaps the first to promote sampling (or what was then called the representative method) over complete enumeration (census), although the oldest reference can be traced back to 1000

BC. In the representative method the sample should mirror the parent finite population and this may be achieved either by balanced sampling on known auxiliary totals through purposive selection or by random sampling leading to equal inclusion probabilities. By the 1920s representative method was widely used. ISI played a vital role by creating an expert committee to report on this method. Bowley's (1926) contribution to the ISI report includes his fundamental work on stratified random sampling with proportional allocation, leading to equal inclusion probabilities. But it was Neyman (1934) who laid the foundations of probability sampling (or design- based approach). He demonstrated that stratified random sampling is preferable to balanced sampling. He also introduced the concept of efficiency and optimal sample allocation that minimizes cost for a specified precision by relaxing Bowley's condition of equal inclusion probabilities. He also showed that for large samples one could obtain confidence intervals on the population mean of a variable of interest such that the frequency of errors in the confidence statement in repeated sampling does not exceed the limit prescribed in advance , 'whatever the unknown properties of the population'. In recent years, balanced sampling, originally advocated by Gini and Galvani, has been refined to incorporate the nice features of both probability sampling and balanced sampling on known auxiliary totals (Deville and Tille 2004). The balanced sampling method is now used in Europe, especially in France, to select samples for establishment surveys.

The 1930s witnessed a rapid growth in demand for socio-economic information, and the advantages of probability sampling in terms of greater scope, reduced cost, greater speed and model-free inferences, were soon recognized worldwide, leading to an increase in number and type of surveys based on probability sampling and covering large populations. Neyman's probability sampling (or design based approach) was almost universally accepted and it became a standard tool for empirical research in social sciences and official statistics. It was also recognized that the precision of an estimator is determined largely by the sample size and not by the sampling fraction. In the early stages of sampling theory development, focus was on estimating totals and means and associated sampling errors, assuming no non-sampling errors such as nonresponse, coverage errors and measurement or response errors.

We now list a few important post-Neyman theoretical developments in the design-based approach. Mahalanobis used multi-stage sampling designs for crop surveys in India as early as 1937. His classic 1944 paper (Mahalanobis 1944) rigorously formulated cost and variance functions for the efficient design of surveys. He was instrumental in creating the National Sample Survey of India, the largest multi-subject continuing survey with full-time staff using personal interviews for socio-economic surveys and physical measurements for crop surveys. 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. Classic text books on sampling by Cochran (1953), Hansen, Hurwitz and Madow (1953), Sukhatme (1954) and Yates (1949) benefited students as well as practitioners greatly.

Survey statisticians at the U. S. Census Bureau, under the leadership of Morris Hansen, made fundamental contributions to sample survey theory and methodology, during the period 1940- 1960. This period is regarded as the golden era of the Census Bureau. Hansen and Hurwitz (1943) developed the basic theory of stratified two-stage cluster sampling with one cluster (or primary sampling unit) within each stratum drawn with probability proportional to size (PPS) and then subsampled at a rate to ensure self-weighting (equal overall probabilities of selection). Unequal probability selection of clusters can lead to significant variance reduction by controlling the variability arising from unequal cluster sizes. Another major contribution from the U. S. Census Bureau is the use of rotation sampling with partial replacement of households to handle response burden in surveys repeated over time, such as the monthly U. S. Current Population Survey for measuring unemployment rates. Hansen et al. (1955) developed simple but efficient composite estimators under rotation sampling. Rotation sampling and composite estimation are widely used in large-scale continuing surveys.

Prior to 1950s, primary focus was on estimating population totals and means. Woodruff (1952) of the U.S. Census Bureau developed a unified approach for constructing confidence intervals on quantiles (in particular median), applicable to general sampling designs.

After the consolidation of the basic design-based sampling theory, Hansen et al. (1951) and others paid attention to measurement or response errors in survey data. Under additive measurement error models with minimal model assumptions on the observed responses treated as random variables, total variance of an estimator is decomposed into sampling variance, simple response variance and correlated response variance (CRV) due to interviewers. The CRV dominates total variance when the number of interviewers is small. Partly for this reason, self-enumeration by mail was first introduced in the 1960 U. S. Census to reduce the CRV component. Earlier, Mahalanobis (1946) developed the method of interpenetrating subsamples for assessing both sampling and interviewer errors. By assigning the subsamples at random to interviewers, both the total variance and the interviewer differences can be assessed. Nonresponse in surveys was also addressed in the early period of survey sampling development. Hansen and Hurwitz (1946) proposed two-phase sampling in which the sample is contacted by mail in the first phase and a subsample of nonrespondents is then subjected to personal interview, assuming complete response or negligible nonresponse at the second phase. This method has been revived recently in Canada after the compulsory long form was replaced by a voluntary National Household Survey.

Attention was also given to inferences for unplanned subpopulations (called domains) such as age-sex groups within a state. Hartley (1959) and Durbin (1958) developed unified theory for domain estimation applicable to general designs and yet requiring only existing formulae for population totals and means.

Most of the survey sampling theory in the early period was developed by official statisticians while academic researchers, especially in USA, did not pay much attention to survey sampling. An exception was, Iowa State University where faculty played a leading role from the early stages under the leadership of Cochran, Jessen and Hartley. Another institution making early contribution to survey practice and research is the Survey Research Center (now Institute for Social Research) at the University of Michigan established 1947, with Leslie Kish as one of its first members.

In the 1950s formal theoretical frameworks for design-based inference on totals and means were proposed by Horvitz and Thompson (1952) and Godambe (1955), by regarding the sample data as the set of sample labels together with the associated variables of interest. Godambe (1955) proposed a general class of linear estimators by letting the sample weight of a unit depend on the label as well as on the labels of the other units in the sample. He then showed that the best linear unbiased estimator does not exist in this general class even under simple random sampling.

2. Inferential Issues: 1950 -

Theoretical foundations

Attempts were made to integrate sample survey theory with mainstream statistical inference via the likelihood function. Godambe (1966) showed that the likelihood function from the full sample data including labels, regarding the vector of unknown population values as the parameter, provides no information on the non-sampled values and hence on the population total or mean. This uninformative feature of the likelihood function is due to the inclusion of labels in the data which makes the sample unique. An alternative design-based route ignores some aspects of the sample data to make the sample nonunique and thus arrive at informative likelihood functions (Hartley and Rao 1968, Royall 1968), This non-parametric likelihood approach is similar to the currently popular empirical likelihood (EL)

approach in mainstream statistical inference (Owen 1988). The EL approach has been applied to sampling problems in recent years to estimate not only totals and means but also more complex parameters. So the integration efforts with main stream statistics was partially successful.

The model-dependent approach provides an alternative route to inference from survey data. The approach requires that the population structure obeys a specified super-population model. The distribution induced by the assumed model provides the basis for inferences. (Brewer 1963, Royall 1970). Such conditional (conditional on the sample) inferences can be appealing. However, the resulting estimators may not be design consistent and they can perform poorly in large samples under model misspecification (Hansen et al. 1983).

A hybrid approach, called model-assisted approach attempts to combine the desirable features of the design-based and model-dependent methods. It entertains only design consistent estimators of the total that are also model unbiased under the assumed working model. This approach is useful for large samples and it leads to valid design-based inferences in large samples, regardless of the validity of the working model. However, efficiency of the estimators does depend on the working model approximately representing the true population structure. Model-assisted estimators are popularly known as generalized regression estimators (GREGs) and are implemented in survey software packages. The GREGs are very attractive in a regular production environment (Brakel and Bethlehem 2008).

Theoretical results for probability based sampling emphasize the first two moments of the sample statistics. An early central limit theorem based on randomization is that of Madow (1948). Hajek(1960) gave a central limit theorem for simple random sampling and a theorem for rejective sampling in (1964). More recent results consider both sequences of fixed finite populations and sequences of finite populations that are samples from a superpopulation,

Variance estimation was very costly, nearly prohibitive, in the 1930's and 1940's, and remains expensive today. Replication was adopted as an efficient method from the beginning. An early form was introduced by Mahalanobis (1939,1946) and called "interpenetrating" samples by him and "random groups" by later authors, see Wolter (2007). The jackknife and bootstrap are the current versions of those early replication procedures. Wolter (2007) credits Durbin (1959) with the first use of the jackknife in finite population estimation. The use of the bootstrap dates from Efron (1979).

Analytic use of survey data

As we have remarked, the early work on probability sampling emphasized totals and means and many estimation procedures were developed for official statistics. However, from the beginning, survey samples were used by social scientists to answer subject matter questions with relevance beyond the finite population sampled. Deming and Stephan (1941) and Deming (1953) gave explicit consideration to the difference between "enumerative" and "analytic" use of survey and census data. The analytic estimates are sometimes called estimates for a superpopulation. Early analysts often treated the sample as a simple random sample and constructed estimates on that basis. Recognizing the potential for bias from ignoring the design, samplers generated considerable research. The theory for analytic estimates developed by survey statisticians has several components. One component are tests for the effect of weights on estimates, see DuMouchel and Duncan (1983), Fuller (1984), and Korn and Graubard (1995). The second component has been the development of design based theory for complicated statistics. See Fuller (1975), Rao and Scott (1981, 1984), and Binder and Roberts (2003). The third approach attempts to build the sampling design into the model. See Skinner (1994) and Pfeffermann and Sverchkov (1999). A number of computer packages (SAS, SUDAAN, R, STATA) are now available that can compute relatively complicated statistics using survey weights.

Many of the algorithms date from the work at Iowa State University (Hidiroglou, Fuller and Hickman, 1976). Another early program was SUDAAN developed by B. V. Shah.

Missing data

Almost all samples (and experiments) have missing or (and) incorrect data. One method of handling missing data is to report the nature and number of missing items and tabulate the remaining items. This was common in the early years, but it was evident that the implied assumption of exchangeability was not reasonable. An early method of correcting for nonresponse was to use a substitute respondent, often interviewing someone "close" to the nonrespondent. A common modification at the analysis stage was, and remains, post stratification. Various forms of imputation in surveys have been used since the beginning, often performed by clerks. An early formal model based imputation was the hot deck imputation procedure used by the U.S. Census Bureau in the 1947 Current Population Survey, see Andridge and Little (2010). Improved computing power and theoretical advances (Kalton and Kish 1984, Rubin 1976, 1987) have made imputation a standard part of estimation for survey samples and an active area of research.

Small area estimation

Demand for estimates for domains whose sampling errors made the direct estimator of little value led to the use of models to construct model based estimates for the domains. Schaible (1996) and Purcell and Kish(1979) give early examples of small area estimation. The U.S. Census Bureau used modelo methods as early as 1947. More recently random models have become important. Early uses of random models for small area estimation are Fay and Harriott (1979) and Battese, Harter and Fuller (1988). Small area models are typically used to produce a small set of estimates for a small number of variables. One can view the typical set of small area estimates as a reallocation of the domain estimates, retaining the direct design-consistent estimate of the grand total. As such, and on the basis of necessity, small area estimates have gained popularity. There has been a large increase in the literature and the field now boasts regular meetings and a book (Rao 2003) with forthcoming second edition, Rao and Molina (2015).

3. The Future

We can project a number of current situations into the future. First, budgets will be tight. Second, demand will grow. There will be demand for estimates for even smaller areas. There will be demand for forecasts. There will be a demand for improved access by users. There will be demand for statistics to be produced even more rapidly. There will be pressure to bring estimates from different sources into agreement.

Bellhouse (2000) discussed the link between developments in survey sampling theory and improvements in computer technology. Citro (2014) and Brick (2011) described how the technological evolution has impacted survey sampling. We can project some of the trends into the future. Computing will become faster, and this will impact all aspects of the field. More complex edit and imputation algorithms will be developed. The time from collection to publication will be shortened. More complex analyses will be performed on survey data. Record linkage procedures will be improved. Data will be made available in different forms. Searchable databases where the user provides queries will become more common.

The use of auxiliary data of all kinds, and in particular administrative data, will increase. Administrative data will be used both as auxiliary data and as the direct estimates for certain items. Citro (2014) gives examples of items where administrative data can be used to replace answers to questions in a questionnaire. The increased computing power will aid in processing administrative

data, and other data (big data such as social media data) so that it becomes more useful as auxiliary data. Porter et al. (2014) used Google trends of Spanish words as functional covariates to estimate state proportions of people speaking Spanish using American Community Survey estimates as dependent variables in small area models.

One of the often quoted advantages of samples relative to censuses is cost. The cost structure has changed with increased computing power and seems destined to continue to change. In the United States the National Land Cover Database is a census of land cover. Of course, it is a census subject to classification error. Classification procedures are expected to improve so that the fraction of agricultural data collected remotely will increase. Data collection agencies will invest more in constructing improved auxiliary data files at the population level. Some data now collected on a sample basis will be collected on a population level.

Due to space and time limitations, we have not explicitly discussed data collection. The way in which data collection procedures have been modified with changing technology is perhaps more obvious than the link to theory. Computer assisted data collection is standard and evolving. The use of geo-location technology can be expected to increase. It is safe to forecast the increased use of remote sensing and remote data collection devices. For example, it would be easy to incorporate the physical data collected by something like the Apple Watch into a health study. Less attractive monitoring devices are currently in use in physical activity surveys.

There is active research in improving the total data collection and production activity (Biemer and Lyberg 2003). Groves and Herringa (2006) proposed tools for actively controlling survey errors and costs that can lead to responsive designs for household surveys. In particular, para data (measurements related to the process of collecting survey data) can be used to monitor field work, to make intervention decisions during data collection and to deal with measurement error, nonresponse and coverage errors (Kreuter 2013).

Survey sampling is an application discipline. Data for human subjects are collected in the world as it is. The social, geographic, cultural, technological world as it is. It is most difficult to forecast how our field will be impacted by social and cultural changes, even in the short run. Will the fact that one must assume that almost all of one's public activity and a great deal of one's private activity has potential of being recorded lead to a more relaxed attitude in responding to questions? Will improved monitoring devices make respondents be more willing to permit their physical activities be monitored? Or will all of the incidental monitoring lead to a reaction against organized data collection? Will increased availability of results based on collected data have a positive or negative effect on data collection efforts? What is the impact of Social Media?

The recent experience is that data collection is becoming more and more difficult. Respondents are facing more and more organized data collection activities. The ubiquitous questionnaire on satisfaction for everything from medical services to tooth paste surely must impact an individual's willingness to respond. Therefore, we see no reason to forecast other than a continuing trend in the difficulty in obtaining cooperation from respondents. Associated with that trend will be increased study of the nature of non-respondents and of non-response.

This discussion makes clear that factors external to our discipline will determine our future activities. We will be required to adapt in data collection, data processing, and data presentation-dissemination.

5. References. References are available from the authors.