UNIVERSITÀ CATTOLICA del Sacro Cuore

DIES ACADEMICUS

Università Cattolica del Sacro Cuore, sede di Piacenza
20 marzo 2013

Impact of Sample Surveys on Social Sciences

J. N. K. Rao

Carleton University, Ottawa, Canada

I feel honored to receive the prestigious *Laurea Honoris Causa* from the Catholic University of the Sacred Heart. I thank the Rector and Professor Maurizio Baussola for their initiative.

My expertise is in the theory and methodology of sample surveys. In this talk I would like to mention some important contributions to sample surveys that have greatly influenced the practice, in particular in obtaining reliable social and economic indicators that are used in making public policy decisions.

Some early landmark contributions

Earliest reference to sampling can be traced back to the Indian epic Mahabharata (1000 BC). According to the story, an Indian king estimated the number of fruits on a tree by examining a single twig (small branch) and his assistant later did a census of the fruits on the tree and was amazed at the accuracy of the sample-based estimate. The king explained that "I of dice possess the science and in numbers thus am skilled". The eminent French Mathematician Laplace estimated the total population of France around 1780s by conducting the enumeration of the population in about 700 communes scattered over the country. However, the Norwegian statistician A. N. Kiaer (1897) is perhaps the first to promote sampling of a finite population (or what he called the "representative method") over complete enumeration (or census). In the representative method the sample should mirror the parent finite population and this may be achieved either by balance sampling through purposive selection or by random sampling. Representative method was used in Russia as early as 1900 and Wright conducted sample surveys in the United States around the same period using this method. By the 1920s, representative method was widely used, and the International Statistical Institute (ISI) played a prominent role by creating a committee in 1924 to report on the representative method. This committee consisted of the British statistician Bowley, the great Italian statistician Corrado Gini and two others. Gini is of course famous for his

income inequality measure, the Gini coefficient, but he has also done fundamental work on sample surveys. Gini believed that majority of methodology work arises from the need to solve concrete practical problems and he considered statistics, economics, sociology, demography and biology as closely related subjects.

The ISI committee's 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. According to this method, the population is divided into homogeneous subgroups (or strata) and random samples are then selected independently from each stratum, using stratum sample sizes proportional to the stratum population sizes. Soon after the 1925 International Statistical Institute meetings in Rome, the Italian Statistical Office wanted to conduct a new census and to make room for the new census data they wanted to keep a "representative sample" of the 1921 census. Gini and Galvani (1929) undertook this task by selecting a purposive sample of 29 administrative units from the 1921 Italian census such that the sample is "balanced" on seven important variables in the sense that the sample mean is close to the population mean for the seven variables.

In 1934 the Polish statistician Jerzy Neyman, famous for his later work on the Neyman-Pearson theory of hypothesis testing, wrote a landmark paper (Neyman 1934)laying the theoretical foundations of probability sampling approach to drawing a sample. He showed that stratified random sampling is preferable to balanced sampling by demonstrating that the Gini-Galvani sample exhibited wide discrepancies with the census counts on some other census variables. 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". Any method of sampling that satisfies the above frequency statement was called "representative" and balanced sampling as formulated by Gini and Galvani is not a member of this class. More recently, balanced sampling of Gini and Galvani has been refined to incorporate the nice features of both probability sampling and balancing on auxiliary variables (Deville and Tille 2004) and the new balanced sampling method is now widely used in Europe, especially in France, to select samples for establishment surveys.

The 1930's saw 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 features, were soon recognized world wide, leading to an increase in the number and type of surveys based on probability sampling and covering large populations. Neyman's probability sampling approach was almost universally accepted and it became a standard tool for empirical research in social sciences and official statistics. It was soon recognized that the precision of an estimator is determined by the *sample size* (number of

units in the sample) and that you do not need to sample a large fraction of the population to get reliable estimators. For example, the precision of an estimator based on a sample of 2000 people is about the same whether the population size is 20 million or a billion.

The great Indian statistician, P. C. Mahalanobis, made pioneering contributions to probability sampling by formulating cost and variance functions for the design of surveys. As early as 1937, he used multi-stage designs for crop yield surveys and he was instrumental in establishing the National Sample Survey of India, the largest multi-subject continuing survey operation with full-time staff using personal interviews for socio-economic surveys and physical measurements for crop surveys. Several prominent Indian survey statisticians were associated with Mahalanobis. Another great Indian statistician, 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. Sukhatme left India in the 1950's to join the Food and Agricultural Organization (FAO) of the United Nations in Rome and promoted sound methods of conducting agricultural surveys world wide.

Survey statisticians at the U.S. Census Bureau, under the leadership of Morris Hansen, 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 introduced stratified two-stage sampling in the context of the U.S. Current Population Survey (which is a monthly survey of labor force characteristics). They selected one primary (or first stage) sampling unit within each stratum with probabilities proportional to size measure (PPS sampling) and then sub-sampled at a rate that ensures approximately equal interviewer work loads which is desirable in terms of field operations. PPS sampling is now widely used in the design of large-scale surveys, but two or more primary units are selected from each stratum to permit the estimation of precision. I made a small contribution to this important topic in the 1960's and two of the methods I have developed for selecting two or more primary units are now used in the monthly Canadian Labor Force Survey and in other surveys, especially in India. Statistics Canada in Canada and ISTAT in Italy played leading roles in developing efficient methods to produce reliable official statistics.

Analysis of complex survey data

The focus of research in sampling theory prior to 1950s was on estimating population totals, means and proportions for the whole population and large planned sub populations, such as states, and associated precisions. For example, a population total is estimated as the weighted sum of the variable of interest for the units in the sample, where the weight is the design weight which can be viewed as the number of population units representing a sample unit. Extensive research has been done on finding efficient estimators of totals and associated measures of precision. Standard text books on sample survey theory provide detailed accounts of estimation of totals and associated precision, and social scientists are familiar with those developments.

Social scientists are also interested in totals and means for unplanned sub populations (called "domains" by the UN Sub Commission on survey sampling, 1947) such as age-sex groups within a state. For example, it would be of interest to study the differences in average income among domains. My Ph.D. supervisor, H. O. Hartley, developed an ingenious method of domain estimation requiring only standard formulas for estimating a population total and published his paper in a special 1959 volume in honor of Corrado Gini. Domain comparisons may be regarded as an example of analysis of survey data.

In practice, social scientists conduct various analyses of survey data, such as regression analysis to study the relationship between a variable of interest and predictor variable or to study the association between two categorical variables. Standard methods of data analysis generally assume that the data are generated by a simple random sample, ignoring the "design effect" due to clustering, stratification, unequal selection probabilities and other design features. However, application of standard methods to survey data, ignoring the design effect, can lead to erroneous inferences even for large samples. In particular, error rates of tests of hypotheses can be much bigger than the nominal levels and the level of stated confidence intervals can be much smaller than the nominal level. Leslie Kish, a famous social survey statistician, 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. In the 1980's I made a small contribution to this important topic by developing suitable corrections to standard tests for categorical data, based on design effect measures that can facilitate secondary analyses from published tables. Those corrections are called Rao-Scott corrections performed well and they are now widely used. Several new software packages for analysis of survey data have incorporated the Rao-Scott corrections. Roberts, Rao and Kumar (1987) developed Rao-Scott type corrections to tests for logistic regression analysis of estimated proportions associated with a binary response variable and applied the methods to a two-way table of employment rates from the Canadian Labor Force Survey 1977 obtained by cross-classifying age and education groups.

Social scientists would like to use micro data files (published or accessed through data resource centers) for analysis of complex survey data. Rapid progress has been made over the past 20 years or so in developing suitable re-sampling methods for analyzing micro-data files. Re-sampling methods select many samples from the given sample repeatedly, in particular the "bootstrap" method I have developed for stratified multi-stage designs (Rao and Wu 1988) which provides bootstrap weights for each bootstrap replicate. All one needs is a data file contained the observed data, the associated weights and the bootstrap weights for each replicate. Software packages that take account of survey weights for estimation of parameters of interest can then be used to calculate correct estimators and associated precisions. As a result, re-sampling methods have attracted the attention of users as they can perform the analyses themselves very easily using standard software packages with weight option. Several recent large-scale surveys at Statistics Canada have adopted the Rao-Wu bootstrap method with 500 bootstrap replicates and users of Statistics Canada survey micro data files seem to be very happy with the bootstrap method for analysis of data. Longitudinal surveys with data on the same individual collected over time are now widely used and

suitable methods for analyses of such data taking the survey design into account have also been developed.

Small area estimation

Traditional design-based methods, inspired by Neyman's work, use domain-specific data and work well when the domain sample sizes are sufficiently large. Such methods, however, may not provide reliable inferences when the domain sample sizes are very small and not implementable if the domains contain no sample units. Domains 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 planning. Clearly, due to cost considerations it is seldom possible to have a large enough overall sample size to support reliable area-specific 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 what is specified at the design stage" (Fuller 1999).

For producing small area estimates with adequate level of precision, it becomes necessary to use indirect methods that borrow information from related small areas through auxiliary information, such as census and current administrative data, to increase the "effective" sample size within the small areas. Realizing the need for indirect estimates, methods that make use of linking models have been proposed in recent years. Success of such model-based methods heavily depends on the availability of good auxiliary information and through model validation. Here social scientists can play an important role in the selection of predictor variables and the form of linking models based on subject matter knowledge. I wrote a book (Rao 2003) on small area estimation giving a comprehensive account of model-based methods, but many important advances have taken place after my book was published and I am now working on the second edition of the book in response to many requests from users.

The "new" methods have been applied successfully worldwide to a variety of small area problems. For example, model-based methods are being used to produce county and school district estimates of poor school-age children in the USA. The U.S. Department of Commerce allocates annually more than $15 billion of funds to counties on the basis of model-based county estimates. The allocated funds support compensatory education programs to meet the needs of educationally disadvantaged children. World wide there is considerable interest in producing reliable small area poverty statistics. Small area estimation is a striking example of the interplay between theory and practice. Practical relevance and theoretical interest of small area estimation have attracted the attention of many researchers, leading to important advances. Italian statisticians have been very active in small area estimation and developed several new methods to estimate poverty rates and other quantities of interest. Professor Maurizio Baussola organized a one day workshop in Piacenza in 2005: "Small area estimation and the local territory" and he edited a special issue of *Rivista*

Internazionale di Scienze Sociali 2008 on small area estimation and I was invited to give the keynote talk at the workshop and also write an overview paper for the special issue.

Current and future research developments

1. Data collection: With the advent of high technology, data collection methods have drastically changed. Current research is focusing on methods than can handle data collection instruments such as cell phones and internet.

2. Hard to reach populations: Lists of units forming the target populations of interest are often not available, for example homeless or illegal immigrant populations. In such cases, multiple incomplete frames containing the units of interest are combined to produce estimates. I have given advice to Professor Fulvia Mecatti of the University of Milan-Bicocca on a simple new method to handle this problem and it performed well in producing reliable estimates using the concept of multiplicity of units observed from the samples drawn from the incomplete frames (Mecatti 2007).

3. Missing data often occurs in socio-economic surveys due to unit and item non-response. Imputation (or filling in missing data) is often used to produce a complete data file so that standard methods for complete data may be applied. However, application of standard methods can lead to erroneous inferences because imputed values are treated as real values. New methods that lead to valid statistical inferences are being developed and this is an active area of research. A related area of interest is statistical matching that can be viewed as a missing data problem where a social scientist desires to undertake joint analyses of variables that are never observed jointly. Aim of statistical matching is to construct a complete data file containing all the variables of interest that can lead to statistically valid analyses.

Concluding remark

I have traced some developments in sample survey theory and methodology that had significant impact on social sciences. Excellent new developments are taking place to address challenging problems in social sciences and other areas as outlined above.

References

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

DeVille, J. C. and Tille, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Fuller, W. A. (1999). Environmental surveys over time. *Journal of Agricultural, Biological and Environmental Statistics*, 4, 3331-345.

Gini, C. and Galvani, L. (1929). Di un'applicazione de metodo rappresentativo all' ultimo censimento Italiano della populazione (1 decembre 1921). *Annali di Statistica*, *Ser VI*, 4, 1-107.

Kiaer, A. N. (1897). Sur les methods representatives ou typologiques appliqués a la statistique. *Bulletin of the International Statistical Institute*m XI, 180-189.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey Methodology*, 33, 151-157.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley, Hoboken, New York.

Rao, J. N. K. and Wu, C. F. J. (1998). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

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