

Catalogue no. 11-522-X
ISSN 1709-8211

Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

Ahead of the Trends: J.N.K. Rao's Contributions to Survey Research

by Sharon L. Lohr

Release date: September 8, 2025



Statistics
Canada

Statistique
Canada

Canada

Ahead of the Trends: J.N.K. Rao's Contributions to Survey Research

Sharon L. Lohr¹

Abstract

J.N.K. Rao has contributed to almost every subdiscipline of survey research, including unequal-probability and two-phase sampling, variance estimation, regression and categorical data analysis, small area estimation, and data integration. For each of these topics, Rao's work anticipated and led future research directions. His contributions will be discussed in the context of broader research trends as seen in the articles of *Survey Methodology* over the journal's 50-year history.

Key Words: Data integration; History of survey research; Replication variance estimation; Survey design; Survey estimation.

1. Introduction

J.N.K. Rao has made numerous fundamental contributions to survey sampling through his publications, talks, consulting with Statistics Canada, and mentorship of young statisticians. Gupta and Mukerjee (2018) provide a bibliography of 279 of Rao's publications from 1956 through early 2018. In this paper, I highlight some of Rao's research in survey design (Section 2), survey estimation (Section 3), and integrating survey data with data from other sources (Section 4), and show how his work has foreseen and spurred subsequent trends in sampling research.

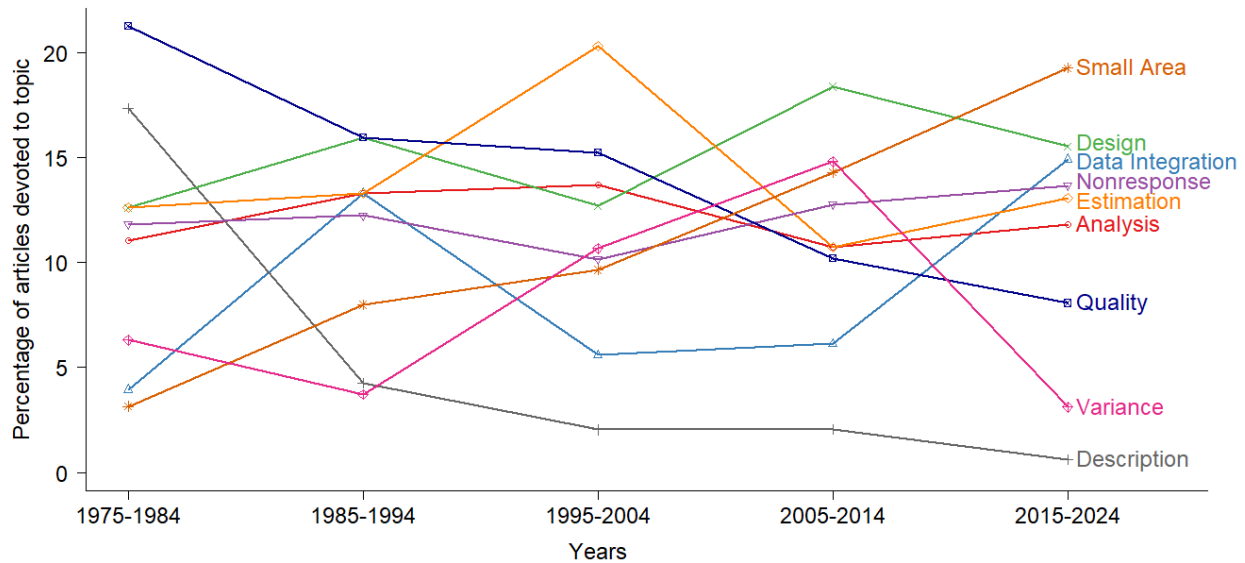
Figure 1-1 illustrates trends in survey research through the relative prevalence of topics in the journal *Survey Methodology*, which celebrates its 50th birthday in 2025. To construct Figure 1-1, I classified each article in *Survey Methodology* from 1975 through June 2024 into one of nine topics: survey description, survey design, estimation of means and totals, analysis of survey data to estimate quantiles or model parameters, variance estimation, methods for treating nonresponse, small area estimation, and data integration. I then divided the articles into ten-year periods and calculated the percentage of articles in each decade that had primary emphasis in each of the nine topics. Of course, this classification is subjective since some papers deal with multiple topics, but the same general trends emerge even with some perturbations of the classification.

Figure 1-1 shows that about 17 percent of the articles in the earliest years of *Survey Methodology* dealt with the **Description** (gray line) of the operation of a particular survey such as the Labour Force Survey, the Canadian Travel Survey, or the Family Expenditure Survey. An additional 21 percent dealt with issues of survey **Quality** (dark blue line) such as measurement error, survey management, meeting user needs, or quality assurance protocols. These two topics plus survey **Design**, in green, accounted for more than half of the articles in the first 10 years of the journal. Survey description and survey quality have represented a smaller percentage of *Survey Methodology* articles over time, as the journal has concentrated more on technical work. The percentage of articles devoted to survey design has remained relatively constant at about 15 percent of the articles.

Articles on **Estimation** of means and totals, in orange, which encompass topics such as weighting and calibration in the full-response setting, ratio and regression estimation, large domain estimation, and foundations of survey inference, have fluctuated between 10 and 20 percent of all articles. **Analysis** papers, in red, deal with topics such as estimating quantiles or regression coefficients, fitting time series models, or performing chi-square tests; these have represented about 10 to 13 percent of all articles. The percentage of articles devoted to **Nonresponse** models, in purple, has also hovered between 10 and 13 percent over each 10-year period. **Variance** estimation (magenta line) accounted for about 15 percent of articles between 2005 and 2014, but only about 3 percent of articles in the most recent 10-year period (although many articles on estimation and analysis methods also involve variance estimation).

¹Sharon L. Lohr, Arizona State University, Box 871804, Tempe, AZ, USA (sharon.lohr@asu.edu)

Figure 1-1
Percentage of *Survey Methodology* articles devoted to each of nine topics by ten-year period, 1975-2024



Two fields have seen substantial increases in their representation in *Survey Methodology*. **Small area** estimation (brown line), which uses models to estimate population quantities for domains with small survey sample sizes, accounted for fewer than 5 percent of articles in the first 10 years and now accounts for nearly 20 percent of all articles. And articles on **Data integration**, which propose or discuss methodology for combining data from different sources (light blue line), are currently undergoing a surge.

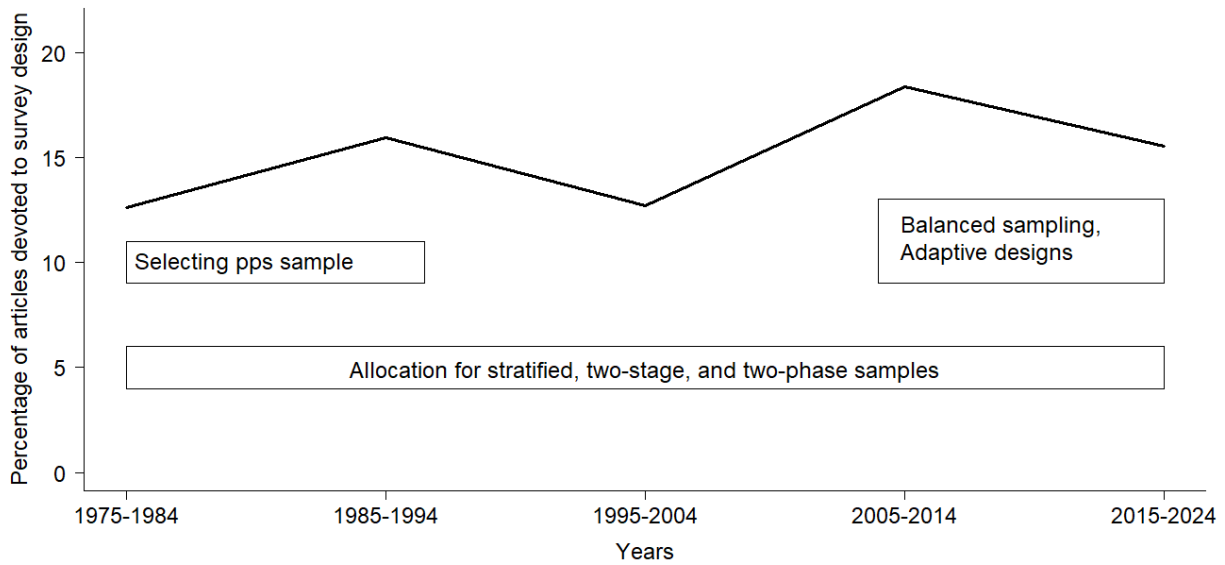
In almost all of these areas, Jon Rao initiated, defined, or foresaw the research trend. In fact, I think this ability to identify what will be important ten or twenty years in the future is one of Rao’s superpowers. I wrote in the preface of the third edition of my sampling textbook that I “have always been awed at his ability to identify and solve the important problems in survey sampling — often years before anyone else realizes how crucial the topics will be” (Lohr, 2022, p. xix).

2. Survey design

The remaining sections of this paper give some examples of Rao identifying important topics before they become a trend. Because of space limitations, I concentrate on his contributions to survey design, estimation, and integrating data. Figure 2-1 shows trends in the area of survey design. During the last 50 years, *Survey Methodology* has published a steady stream of articles on sample allocation for stratified, multistage, and two-phase designs. In recent years, there have been more articles on topics such as balanced sampling and adaptive designs. At the beginning, however, most of the design articles in the journal focused on methods for selecting a probability proportional to size (pps) sample.

Rao’s work in survey design predates the journal’s establishment, with three highly influential papers in 1962 and 1963. Theory for computing estimates and variances for without-replacement pps samples had been developed in the early 1950s (Narain, 1951; Horvitz and Thompson, 1952), but that theory did not say how to operationalize these methods and calculate the joint inclusion probabilities for a specific design. Hartley and Rao (1962) found the joint inclusion probabilities, and hence the variance, when a sample is selected by arranging the clusters in random order and taking a random systematic sample of elements. Rao, Hartley, and Cochran (1962) specified taking a pps sample by dividing the population units into n random groups and then selecting one unit with pps from each group. This method had simple calculations and a variance estimator that was guaranteed to be non-negative. Rao (1963) examined properties of three methods for selecting a pps sample, including a rejective method that draws with-replacement samples until one with all distinct elements is obtained.

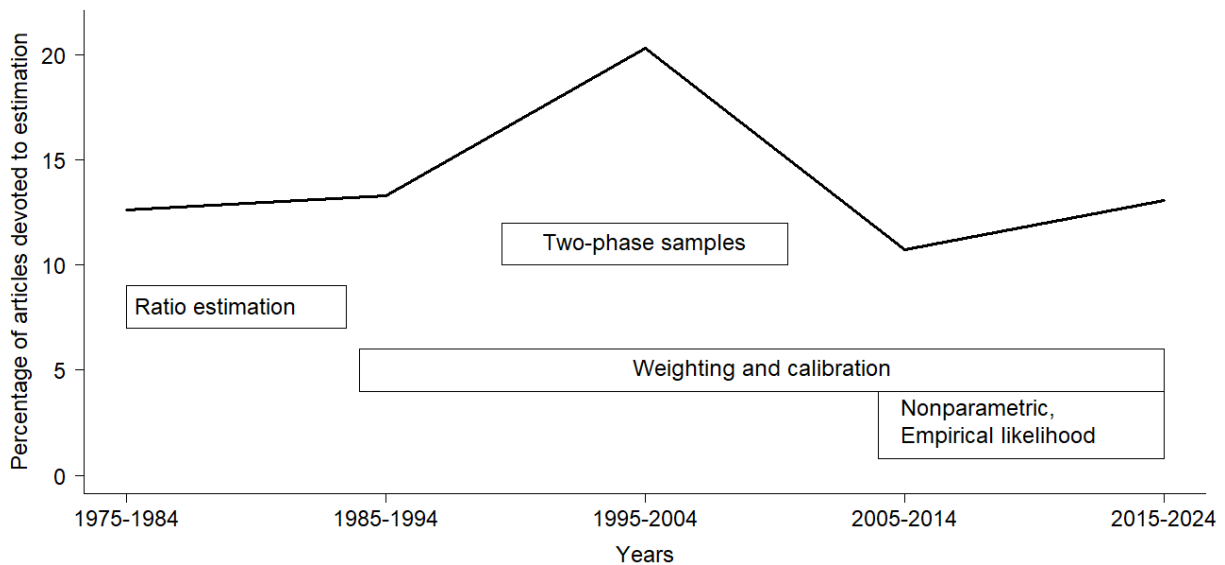
Figure 2-1
Percentage of *Survey Methodology* articles devoted to survey design, 1975-2024



3. Estimation of means, totals, model parameters, and variances

Now let's look at estimation of population means and totals (Figure 3-1). Many of the estimation papers in the early years of *Survey Methodology* dealt with ratio estimation. Later on, large numbers of estimation papers focused on weighting and calibration methods and on estimating population means and totals from two-phase samples. In recent years there has been a focus on empirical likelihood and nonparametric methods. In each of these areas, Rao published influential papers long before the trend took off in *Survey Methodology*.

Figure 3-1
Percentage of *Survey Methodology* articles devoted to estimation of population means and totals, 1975-2024



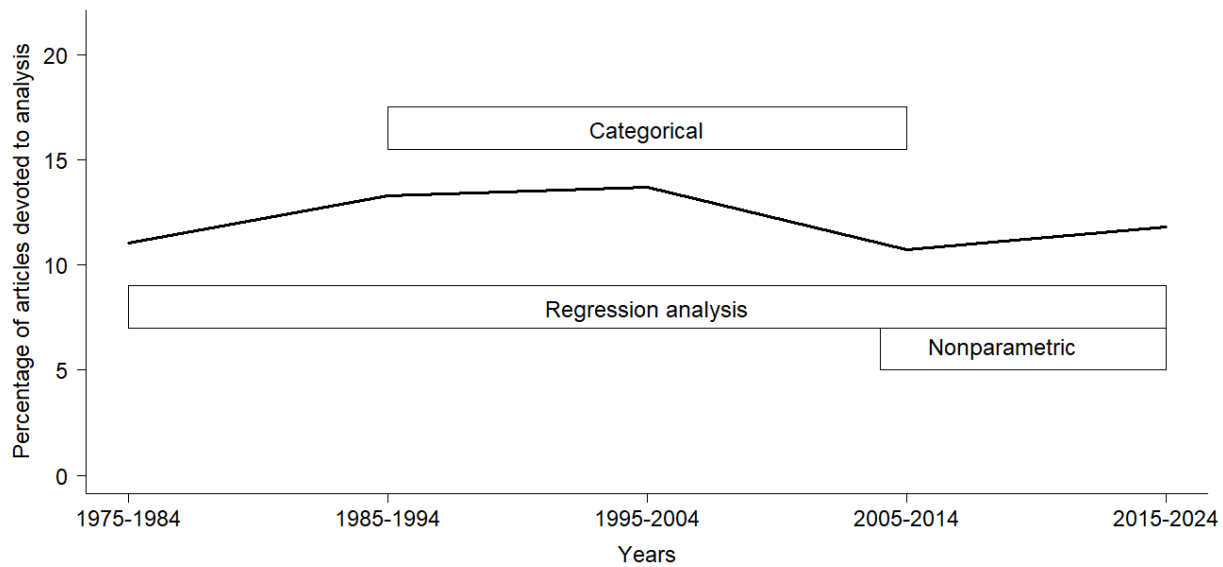
Rao's very first publication dealt with ratio estimation — he looked at the efficiency gains that could be achieved by stratifying the sample and using a combined ratio estimator instead of an unstratified ratio estimator (Rao and Chawla, 1956). He proceeded to write many other papers on small sample properties, variance estimation, comparing separate and combined ratio estimators, and other topics related to ratio estimation.

Rao has been one of the most influential researchers on other estimation topics as well. Rao (1973), which derived properties of two-phase estimators for stratification, preceded the establishment of *Survey Methodology* and the surge of papers in the 1990s and 2000s on two-phase sampling. Regarding nonparametric approaches to survey inference, the nonparametric scale-load approach introduced by Hartley and Rao (1968) anticipated later developments in empirical likelihood methods by about 20 years (see Rao and Wu, 2009, for a review).

Of course, Rao has written extensively on weighting and calibration and making use of auxiliary information. The contribution I want to highlight here is Rao (1994), which consolidated results in the literature and studied optimal regression estimators and conditional inference.

Moving on to estimation of quantities other than means and totals (secondary data analysis), *Survey Methodology* has published numerous papers on estimating quantiles and regression coefficients, and on categorical data analysis with survey data (Figure 3-2). Here again, Rao established the trends in multiple areas. The famous Rao and Scott (1981, 1984) corrections for chi-square tests gave simple adjustments that accounted for the survey design; Scott (2007) discussed the immense impact of this work.

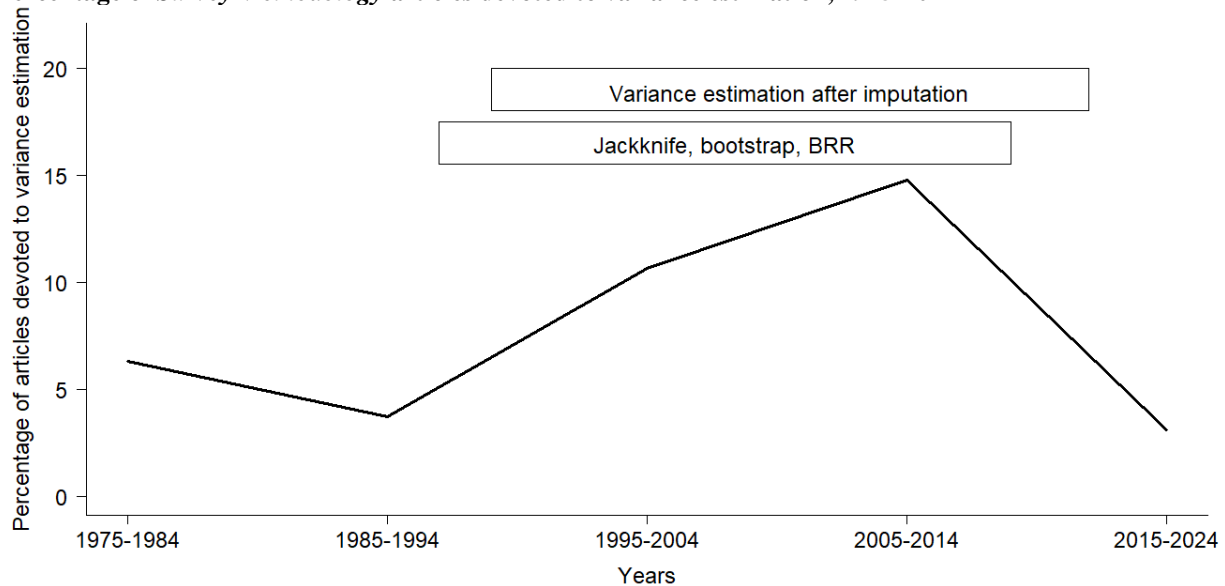
Figure 3-2
Percentage of *Survey Methodology* articles devoted to secondary analysis of survey data, 1975-2024



I just want to mention a couple of Rao's many contributions to regression analysis with survey data. First, Roberts, Rao, and Kumar (1987), based on his student Georgia Roberts's dissertation, set out theory for performing logistic regression with survey data. The second contribution I want to highlight is Rao's work on a weighted composite likelihood method for estimating regression parameters and variance components in a hierarchical model (Rao, Verret, and Hidioglou, 2013; Yi, Rao, and Li, 2016). The estimates of variance components depend on the first-stage inclusion probabilities and on the joint inclusion probabilities for pairs of elements within clusters.

Of course, every estimate needs to be accompanied by a measure of uncertainty, and between 1995 and 2014 about 12 percent of the articles in *Survey Methodology* had primary focus on variance estimation. The increase in activity in these years was primarily due to an influx of articles on resampling methods and on estimating variances after imputation (Figure 3-3).

Figure 3-3
Percentage of *Survey Methodology* articles devoted to variance estimation, 1975-2024



For both of these areas, Rao’s work anticipated the trend. His groundbreaking papers showing the asymptotic properties of jackknife, balanced repeated replication (BRR), and bootstrap variance estimators for surveys preceded the surge of research in these areas (Krewski and Rao, 1981; Rao and Wu, 1985, 1988). The bootstrap replicate weight method in Rao, Wu, and Yue (1990) is used by survey organizations around the world.

A similar surge in research occurred after publication of the Rao and Shao (1992) method of jackknife variance estimation with hot deck imputation and after subsequent work with students Wesley Yung and David Haziza on inference when data are imputed (Yung and Rao, 2000; Haziza and Rao, 2006).

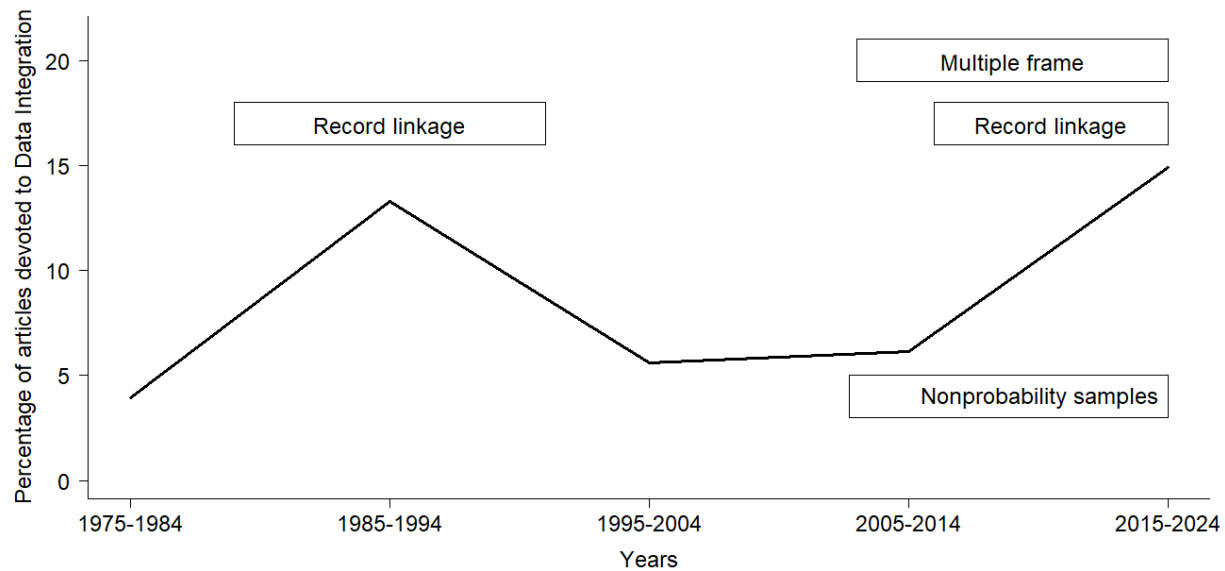
4. Data integration

Data integration encompasses many topics. Poststratification and calibration, which obtain control totals from an external data source, can be thought of as a form of data integration.

Small area estimation is also a data integration method. Composite small area estimators use auxiliary information from another source to obtain regression predictions for individuals or domain means in domains with small sample sizes, and then those predictions are combined with direct domain estimators from the survey. Rao has been highly influential in small area estimation, with more than 60 papers and books on this subject. Prasad and Rao (1990), which established the mean squared error of the Fay-Herriot estimator, has been cited more than 1,000 times in the literature and preceded the increased activity during the 1990s and 2000s seen in Figure 1-1. Rao’s influential book on small area estimation (Rao, 2003, with second edition Rao and Molina, 2015) is cited in almost every paper that has subsequently been published in the now-large area of small area estimation (the area currently accounts for the highest percentage of articles in *Survey Methodology*).

For other areas of data integration (Figure 4-1), there are two periods of intense activity in *Survey Methodology*. The early mode from 1985 to 1994 focused mainly on record linkage techniques and there has been a resurgence of activity in linkage in recent years. But at the present time, as surveys become more expensive and have ever-lower response rates, and as other data sources are increasingly available, there is growing interest in other methods for data integration such as multiple frame sampling, imputation, and modeling.

Figure 4-1
Percentage of *Survey Methodology* articles devoted to data integration, 1975-2024



Rao, of course, published several papers on multiple frame methods that preceded the increased activity in the area. He had long been involved in dual frame surveys through discussions with his advisor H. O. Hartley, and highlighted dual frame surveys in his tribute to Hartley’s contributions to sample survey theory and methods (Rao, 1983). Skinner and Rao (1996), on pseudo-maximum-likelihood estimation for dual frame surveys, solved an important problem for how to combine independent samples in a way that minimizes the variances of estimators yet uses the same set of weights for all variables y .

In the arena of model-based data integration, I want to highlight Kim and Rao (2012), which studies methods for combining a large sample that measures auxiliary information x with a small sample that measures both x and the variable of interest y . In this research, both samples are probability samples, but there have been many extensions involving nonprobability samples. I recommend that anyone interested in the topic of data integration or using nonprobability samples read Rao (2021), which gives his thoughts on integrating probability and nonprobability samples. In this one paper, Rao reviews the history of probability sampling, calibration, small area estimation, and dual frame sampling, and then looks at methods for combining information from a nonprobability sample with information from a probability sample. Rao examines estimators when y is observed in both samples and when y is observed only in the nonprobability sample. The main concern, of course, is selection bias in the nonprobability sample and Rao emphasizes the need for rich auxiliary information that can account for sample selection bias or predict y for population members not in the nonprobability sample.

The theme of this conference is “Shaping the Future of Official Statistics,” and I have discussed how Rao’s groundbreaking work has shaped the present of official statistics. One might naturally ask what Rao is working on now, since, if past patterns continue, those topics may well define the research trends of the next few years. Rao and Fuller (2017) projected some current trends into the future: to respond to demand for faster and more granular statistics (with no compromise of quality), to bring estimates from different sources into agreement, and to make use of ever-improving computational capabilities. Rao and Lohr (2025) summarized recent developments in survey design and data collection, estimation of means and totals, secondary data analysis, variance estimation, nonresponse adjustments, small area estimation, and data integration. Much of Rao’s research in the past few years has been on these problems, and I think it is quite likely that someone looking at his contributions at a conference in the year 2035 would be able to trace the influence of his recent papers on the field of survey sampling.

References

- Gupta, V. K., and Mukerjee, R. (2018), "J. N. K. Rao — A tribute", *Statistics and Applications*, 18, pp. 1-19.
- Hartley, H. O., and Rao, J. N. K. (1962), "Sampling with unequal probabilities and without replacement", *Annals of Mathematical Statistics*, 33, pp. 350-374.
- Hartley, H. O., and Rao, J. N. K. (1968), "A new estimation theory for sample surveys", *Biometrika*, 55, pp. 547-556.
- Haziza, D., and Rao, J. N. K. (2006), "A nonresponse model approach to inference under imputation for missing survey data", *Survey Methodology*, 32, pp. 53-64.
- 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, pp. 663-685.
- Kim, J. K., and Rao, J. N. K. (2012), "Combining data from two independent surveys: A model-assisted approach", *Biometrika*, 99, pp. 85-100.
- 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, pp. 1010-1019.
- Lohr, S. L. (2022), *Sampling: Design and Analysis*, 3rd ed., Boca Raton, FL: CRC Press.
- Narain, R. D. (1951), "On sampling without replacement with varying probabilities", *Journal of the Indian Society of Agricultural Statistics*, 3, pp. 169-174.
- Prasad, N. G. N., and Rao, J. N. K. (1990), "The estimation of mean squared errors of small area estimators", *Journal of the American Statistical Association*, 85, pp. 163-171.
- Rao, J. N. K. (1963), "On three procedures of unequal probability sampling without replacement", *Journal of the American Statistical Association*, 58, pp. 202-215.
- Rao, J. N. K. (1973), "On double sampling for stratification and analytical surveys", *Biometrika*, 60, pp. 125-133.
- Rao, J. N. K. (1983), "H. O. Hartley's contributions to sample survey theory and methods", *The American Statistician*, 37, pp. 344-350.
- Rao, J. N. K. (1994), "Estimating totals and distribution functions using auxiliary information at the estimation stage", *Journal of Official Statistics*, 10, pp. 53-165.
- Rao, J. N. K. (2003), *Small Area Estimation*, New York: Wiley.
- Rao, J. N. K. (2021), "On making valid inferences by integrating data from surveys and other sources", *Sankhyā B*, 83, pp. 242-272.
- Rao, J. N. K. and Chawla, H. K. (1956), "Efficiency of stratification in sub-sampling designs for the ratio method of estimation with varying probabilities of selection", *Journal of the Indian Society of Agricultural Statistics*, 8, pp. 91-101.
- Rao, J. N. K. and Fuller, W. A. (2017), "Sample survey theory and methods: Past, present, and future", *Survey Methodology*, 43, pp. 145-160.
- 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, pp. 482-491.

- Rao, J. N. K. and Lohr, S. L. (2025), "Trends and directions in sample survey theory and methods", to appear, *Survey Methodology*.
- Rao, J. N. K. and Molina, I. (2015), *Small Area Estimation, 2nd Ed.* Hoboken, NJ: Wiley.
- 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, pp. 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", *Annals of Statistics*, 12, pp. 46-60.
- Rao, J. N. K. and Shao, J. (1992), "Jackknife variance estimation with survey data under hot deck imputation", *Biometrika*, 79, pp. 811-822.
- Rao, J. N. K., Verret, F., and Hidroglou, M. (2013), "A weighted composite likelihood approach to inference for two-level models from survey data", *Survey Methodology*, 39, pp. 263-282.
- Rao, J. N. K. and Wu, C. F. J. (1985), "Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics", *Journal of the American Statistical Association*, 80, pp. 620-630.
- Rao, J. N. K. and Wu, C. F. J. (1988), "Resampling inference with complex survey data", *Journal of the American Statistical Association*, 83, pp. 231-241.
- Rao, J. N. K., Wu, C. F. J., and Yue, K. (1990), "Some recent work on re-sampling methods for complex surveys", *Survey Methodology*, 18, pp. 209-217.
- Rao, J. N. K. and Wu, C. (2009), "Empirical likelihood methods", in D. Pfeiffermann and C. R. Rao (eds.) *Sample Surveys: Inference and Analysis, Vol. 29B*, Amsterdam: Elsevier, pp. 189-207.
- Roberts, G., Rao, J. N. K., and Kumar, S. (1987), "Logistic regression analysis of sample survey data", *Biometrika*, 74, pp. 1-12.
- Scott, A. J. (2007), "Rao-Scott corrections and their impact," *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 3514-3518.
- Skinner, C. J., and Rao, J. N. K. (1996), "Estimation in dual frame surveys with complex designs", *Journal of the American Statistical Association*, 91, pp. 349-356.
- Yi, G., Rao, J. N. K., and Li, H. (2016), "A weighted composite likelihood approach for analysis of survey data under two-level models", *Statistica Sinica*, 26, pp. 569-587.
- Yung, W., and Rao, J. N. K. (2000), "Jackknife variance estimation under imputation for estimators using poststratification information", *Journal of the American Statistical Association*, 95, pp. 23-31.