The Survey Statistician, 2023, Vol. 88, 51–57.

Survey Sampling History at Iowa State University

Jae Kwang Kim

Iowa State University, USA, jkim@iastate.edu

Abstract

Iowa State University (ISU) has played an important role in research and education in survey sampling. In this article, we give a brief history of the Statistical Laboratory Survey Section (now known as the Center for Survey Statistics and Methodology) and its impact on the survey sampling community. Some reflections on my journey in survey sampling are also presented.

Keywords: Center for Survey Statistics and Methodology, National Resources Inventory.

1 Introduction

Iowa State University (ISU) has played an important role in research, practice, and education in survey sampling. The Statistical Laboratory (Stat Lab) at Iowa State College was established in 1934 to promote statistical research and provide consulting to other university units, and led by George Snedecor. The Survey Section of the Statistical Laboratory, which later became the Center for Survey Statistics and Methodology (CSSM), was established in 1938 as a result of a cooperative agreement between the Statistical Laboratory and the U.S. Department of Agriculture (USDA).

In this article, I give a brief history of Stat Lab and its impact on the survey sampling community. Some reflections on my journey in survey sampling are also presented.

2 History

2.1 Early Years, 1938-1948

The Department of Agriculture was one of the early organizations in the U. S. to initiate research and development work on probability sampling, and they established a cooperative research program with the Statistical Laboratory at Iowa State University in 1938. Initial work under the cooperative agreement with the USDA led to the development of the Master Sample of Agriculture (King and Jessen, 1945), a national area sample of land that was subsequently used in numerous economic surveys of American agriculture, as witnessed by Fuller (1984).

Jessen (1942) investigated the problem of approximating the optimum sizes of sampling units for agricultural studies. The paper guided the development of the Master Sample of Agriculture and stimulated the later development of designs and theory for rotating samples for surveys taken on successive occasions for time series estimation.

Iowa State is also one of the few universities that has featured training in sample survey theory and methods as an important part of statistical training, as noted by Hansen (1987). Under George Snedecor, Iowa State was also a center for statistical treatment of experimental work. The emphasis in applied statistics at Iowa was then on sample surveys and experimental design. William G. Cochran then joined the Rothamsted Experimental Station and lectured on sample surveys and experimental design. *Sampling Techniques* by Cochran (1953) was developed from the lectures at Iowa State University. In 1946, Cochran left Iowa to organize and head the graduate program in experimental statistics at North Carolina State College at Raleigh. Theodore Bancroft took over from George Snedecor as the Head of the Department of Statistics (established in 1947) and the Director of the Statistical Laboratory during 1950-1972.

2.2 1950s-1960s

Survey research and consulting continued to grow through the efforts of the groups in many disciplines and areas. The early 1950s saw survey statistics research highlighted in Ph. D. dissertations in other disciplines on campus, a practice that has grown significantly since. Other efforts in the 1950s and 1960s included survey research, methodology, and practice in areas such as employment trends in the state of Iowa, home economics studies across the nation, farm practice surveys, and surveys for researching social welfare in the state of Iowa, just to name a few. Interest from international venues also increased after the 1950s, as research fellows from across the globe began visiting the department to take courses in survey methodology and work with faculty and students on applied research projects. For example, the seminal paper of Horvitz and Thompson (1952) was published when the authors were graduate students at ISU and they were influenced by the lectures from Midzuno who visited ISU from Japan. P. V. Sukhatme also visited from India and wrote his famous book (Sukhatme, 1953) at ISU, which was published by Iowa State College Press. His brother, B. V. Sukhatme, joined as a faculty member of ISU later and revised the book (Sukhatme and Sukhatme, 1970) and advised graduate students.

In 1956, the Survey Section began cooperating with the US Soil Conservation Service (now the USDA Natural Resources Conservation Service) to develop survey methods and provide operational support for the National Resources Inventory (NRI), a longitudinal survey of agricultural and other natural resources on nonfederal lands. CSSM continues to work on the NRI project today, and the survey has been the inspiration for many methodologies related to sampling and estimation (Nusser and Goebel, 1997). Results from these surveys are used extensively in the construction of Farm Bills in the US. A second major, ongoing collaboration began in this period with the US Census Bureau on improving survey sampling and methodology in census practices.

H. O. Hartley joined ISU in 1953 and became deeply involved in research and teaching. Hartley made several major contributions, including the famous paper on unbiased ratio estimation published in Nature (Hartley and Ross, 1954). This paper motivated Mickey to develop a whole class of unbiased regression estimators (Mickey, 1959), which was done at ISU. Hartley also wrote seminal papers on domain estimation (Hartley, 1959) and dual frame surveys (Hartley, 1962) while he was at Iowa State. Hartley's contribution to survey sampling is well summarized by Rao (1983). Hartley is the Ph. D. advisor of J. N. K. Rao. Rao stayed at ISU for 5 years (1958-1963), three years as a student and two years as Assistant Professor (AP). Rao, Hartley, and Cochran (1962) published a paper on a very simple procedure of unequal probability sampling scheme without replacement that allows them to estimate the variance of the resulting estimator of total. During his AP tenure, Rao shared an office with Wayne Fuller, who just joined the Statistics Department as an AP, and they remained good friends to each other throughout their professional careers. Hartley also advised Edward Bryant on two-way stratification (Bryant et al, 1960), which received a lot of attention at that time. Edward Bryant later founded Westat.

Figure 1: Wayne A. Fuller

2.3 1970s-1980s

With the arrival of more powerful mainframe computers, namely the IBM 360, as well as advances in statistical computing, the Survey Section began to develop and implement more sophisticated survey and data analysis. SUPER CARP, the first mainframe computer program developed by Mike Hidiroglou in the Survey Section, allowed the implementation of many estimation methods used in survey sampling in an automated manner. The program used the software developed to compute regression estimation in the context of survey sampling. The software was also expanded to allow for estimation and estimated standard errors for totals, ratios, means, and proportions for subdivisions of a sampled population. It also contained several procedures appropriate for data observed subject to measurement errors. It was operational by 1976: See Hidiroglou, Fuller, and Hickman (1978). In 1985, SUPER CARP underwent major revisions and updates of algorithms that allowed its deployment on IBM-PCs. Later, EV (errors in variables) CARP was released as companion software to Wayne Fuller's book, *Measurement Error Models* (1987). Many graduate students, including Cary Isaki, Mike Hidiroglou, Kirk Wolter, Elizabeth Huang, Yasuo Amemiya, Sastry Pantula, David Dickey, and John Eltinge, worked on this project and its related topics under the supervision of Wayne Fuller. Kirk Wolter served as the president of the International Association of Survey Statisticians (IASS) during 1999-2001.

Wayne Fuller has since made many important contributions to survey sampling. Wayne Fuller can be credited for introducing the regression idea to adjust the design weights to construct calibration weights. Huang and Fuller (1978) developed an iterative method for constructing range-restricted weights that meet the benchmarking constraints and the design consistency. Isaki and Fuller (1982) laid the foundation for establishing the optimality of the regression estimator. Battese, Harter, and Fuller (1988) developed a framework for a unit-level model approach to small area estimation.

2.4 1990s-2000s

New faculty members, including Sarah Nusser, F. Jay Breidt, and Jean Opsomer, joined the Survey Section in the 1990s. During this period, survey statistics education and research continued to flourish at CSSM. The wide diversity of projects undertaken included survey consulting on projects such as local efforts to analyze and improve ISU campus services; consulting projects with departments and bureaus of the state of Iowa to develop statistical pictures and gather information about Iowa residents' behaviors and preferences to farm production surveys and Iowa business and economic survey research; ongoing long-term research projects with national agencies such as the USDA / NCRS and the Census Bureau, as well as new projects with the National Cancer Institute, the Centers for Disease Control, the National Institutes of Health, the Bureau of Land Management, and the

National Science Foundation, amongst many others. In November 2002, the Survey Section of the Stat Lab officially became the Center for Survey Statistics and Methodology. Wayne Fuller officially retired in 2001, but he continued working as a part-time consultant at the CSSM.

On the research side, a measurement error model was applied to estimate the usual daily intake distribution (Nusser Carriquiry, Dodd, and Fuller, 1996). This work served as the basis of ongoing work by Alicia Carriquiry and Fuller that still influences approaches to dietary assessment in the United States and many other countries. The NRI's longitudinal 2-stage stratified cluster sample, which is observed every 5 years, was redesigned to an annual survey with supplemented panels. Design and estimation of the supplemented panel survey became very important for NRI application (Nusser, Breidt and Fuller, 1998; Fuller, 2003). Breidt and Opsomer (2000) developed nonparametric regression estimation methods and collaborated on many research problems in survey sampling. Regression weighting methods were further developed for the U. S. Census (e. g., Isaki, Tsay, and Fuller, 2004). Fractional hot deck imputation was developed by Kim and Fuller (2004). Emily Berg wrote a Ph. D. thesis on small area estimation under the supervision of Wayne Fuller (Berg and Fuller, 2014). An advanced-level textbook on survey sampling written by Fuller (2009) was finally published.

2.5 2010s-present

Cindy Yu, Jae Kwang Kim, Zhengyuan Zhu, and Emily Berg joined ISU and became a new generation of CSSM faculty. Each has influenced continuous methodological developments in different ways - Kim via missing data analysis, Zhu via spatial data models, and Berg via her expertise in small area estimation. CSSM provided statistical consulting to other agencies, including the National Agricultural Statistical Service, the Bureau of Land Management, the Bureau of Justice Statistics, and the Food and Agriculture Organization (FAO) of the United Nations.

Kim expanded the departmental curriculum with a graduate-level course on handling missing data, and the lecture notes matured into a textbook (Kim and Shao, 2021). Kim used his expertise in missing data and developed a series of methods for data integration (Kim, 2022). Under the directorship of Zhu, the CSSM continues to expand, and the funding size is now about 5 million USD per year.

3 Reflections

I consider myself to belong to the third generation in the survey sampling community. The first generation at ISU includes William Cochran and H. O. Hartley. The second generation at ISU includes J. N. K. Rao and Wayne Fuller. Standing on the shoulders of giants, I learned the essence of survey sampling theory and methods. When I was a graduate student at ISU, the Statistics Department offered a very specialized curriculum in survey sampling, with separate MS and Ph. D. courses in sampling, and Jay Breidt's lectures were very clear and excellent. I wrote my dissertation under the supervision of Wayne Fuller and have benefited a lot from Fuller's excellent insights and rich research experience.

The second generation flourished in the "Golden Age of Survey Research" (Singer, 2016) when the response rates were high and other data sources were unavailable. As a third-generation member, as Kalton (2019) pointed out, I faced two main challenges in survey sampling. One is the declining trend in response rates and the related increases in the costs of surveys based on probability samples. The other challenge comes from the emergence of an alternative source of information, including large administrative data and low-cost web panel samples. Thus, naturally, I became interested in the research topics addressing these new challenges: handling missing data and adjusting selection bias in the voluntary samples through data integration or weighting.

Imputation for handling item nonresponse is a topic of my Ph. D. thesis. I worked on a consulting

project for the U. S. Bureau of Census on estimating the variance of the census long-form survey estimates after nearest neighbor imputation (Kim, Fuller, and Bell, 2011), which is based on ignorable missingness assumption. Before joining ISU in 2008, I had the opportunity to work on a project related to election exit polls in Korea. This sparked my interest in nonignorable missing research and led to several papers over the years at Iowa (Kim and Yu, 2012; Morikawa and Kim, 2021). An invitation from J. N. K. Rao to visit Ottawa in 2007 was also an eye-opening experience for me. Combining information from two independent surveys (Kim and Rao, 2012) started with the visit to Rao. Consulting projects from Statistics Korea and the USDA National Agricultural Statistics Service, data integration methods were developed using the measurement error model (Kim, Park, and Kim, 2015) and multilevel models (Kim, Wang, Zhu, and Cruze, 2018), respectively. A visit to the Australian Bureau of Statistics in 2016 sparked my interest in data integration research incorporating big data (Kim and Tam, 2021).

As a sampling statistician in academia, I now see another challenge approaching us: how to teach survey sampling and educate the next generation so that they can understand the value of survey sampling. In the era of machine learning and AI, students are more interested in learning modern techniques than classical subjects. Thus, in addition to the decline in survey participation, we are facing a decline in interest in survey sampling research among the next generation. How do we improve our teaching, modernize our textbook and find interesting research problems to attract young "smart" students into survey sampling? I think these questions should be seriously addressed by the survey sampling community in academia.

Acknowledgements

I would like to thank Danutė Krapavickaitė for the invitation to write an article on the history of survey sampling at ISU. I also thank Jon Rao, Sarah Nusser, Mike Hidiroglou, and Wayne Fuller for their very helpful comments on the earlier version of the article. Some contents are taken from the website of the CSSM. The research was partially supported by the cooperative agreement between NRCS-USDA and the CSSM at ISU and the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa.

References

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **83**, 28–36.

Berg, E. J. and Fuller, W. A. (2014). Small Area Prediction of Proportions with Applications to the Canadian Labour Force Survey. *Journal of Survey Statistics and Methodology* **2**, 227–256.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics* **28**, 1026–1053.

Bryant, E. C., Hartley, H. O., and Jessen, R. J. (1960). Design and Estimation in two-way stratification. *Journal of the American Statistical Association* **55**, 105–124.

Cochran, W. G. (1953). *Sampling Techniques*. John Wiley & Sons, New York.

Fuller, W. A. (1984). The Master Sample of Agriculture. In David, H. A. and David, H. T., Eds. *Statistics: An Appraisal*, Iowa State University Press, Ames, Iowa, 583–602.

Fuller, W. A. (2003). Estimation for multiple phase samples. In *Analysis of Survey Data* (edited by R.L. Chambers and C.J. Skinner), Wiley, Chichester, UK, 307–322.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons, New York.

Fuller, W. A. (2009). *Sampling Statistics*. John Wiley & Sons, Hoboken, New Jersey.

Hansen, M. H. (1987). Some History and Reminiscences on Survey Sampling. *Statistical Science* **2**, 180–190.

Hartley, H. O. (1959). Analytical Studies of Survey Data, *Rome: Istituto di Statistica. Volume in Honor of Corrado Gini*. pp. 1–32.

Hartley, H. O. (1962). Multiple Frame Surveys, *Proceedings of Social Statistics Section*. pp.203–206. American Statistical Association.

Hartley, H. 0., and Ross, A. (1954). Unbiased ratio estimators. *Nature* **174**,270-271.

Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. (1978). SUPER CARP. Statistical Laboratory, Iowa State University, Ames, Iowa.

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 Section on Survey Research Methods*, American Statistical Association, 1897–1904.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association* **77**, 89–96.

Isaki, C. T., Tsay, J. H., and Fuller, W. A. (2004). Weighting sample data subject to independent controls. *Survey Methodology* **30**, 35–44.

Jessen, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin* 304.

Kalton, G. (2019). Developments in Survey Research over the Past 60 years: A personal perspective. *International Statistical Review* **87**, S10-S30.

Kim, J. K. (2022). A gentle introduction to data integration in survey sampling. *The Survey Statistician* **85**, 19–29.

Kim, J. K. and Fuller, W. A. (2004) Fractional hot deck imputation. *Biometrika* **91**, 559–578.

Kim, J. K., Fuller, W. A., and Bell, W. R. (2011) Variance Estimation for Nearest Neighbor Imputation for U.S. Census Long Form Data, *Annals of Applied Statistics* **5**, 824–842.

Kim, J. K., Park, S. and Kim, S. (2015). Small area estimation combining information from several sources, *Survey Methodology* **41**, 21–36.

Kim, J. K. and Rao, J. N. K. (2012). Combining data from two independent surveys: a model-assisted approach, *Biometrika* **99**, 85–100.

Kim, J. K. and Shao, J. (2021). *Statistical Methods for Handling Incomplete Data* (2nd Edn). Chapman & Hall / CRC, Boca Raton, FL.

Kim, J. K. and Tam, S. (2021). Data integration by combining big data and survey sample data for finite population inference, *International Statistical Review* **89**, 382–401.

Kim, J. K. and Yu, C. L. (2011). A semi-parametric estimation of mean functionals with non-ignorable missing data, *Journal of the American Statistical Association* **106**, 157–165.

Kim, J. K., Wang, Z,, Zhu, Z., and Cruze, N. (2018). Combining survey and non-survey big data for improved sub-area prediction using a multi-level model, *Journal of Agricultural, Biological, and Environmental Statistics* **23**, 175–189.

King, A. J. and Jessen, R. J. (1945). The master sample of agriculture. *Journal of the American Statistical Association* **40**, 38–56.

Mickey, M. R. (1959). Some finite population unbiased ratio and regression estimator. *Journal of the American Statistical Association* **54**, 594–612.

Morikawa, K. and Kim, J. K. (2021). Semiparametric Optimal Estimation With Nonignorable Nonresponse Data, *Annals of Statistics* **49**, 2991-3014

Nusser, S. M., Breidt, F. J., and Fuller, W. A. (1998). Design and estimation for investigating the dynamics of natural resources. *Ecological Applications* **8**, 234–245.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996). A semiparametric transformation approach to estimating usual daily intake distribution. *Journal of the American Statistical Association* **91**, 1440–1449.

Nusser, S. M. and Goebel, J. J. (1997). The National Resources Inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics* **4**, 181–204.

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. (1983). H.O. Hartley's Contributions to Sample Survey Theory and Methods. *The American Statistician*, **37**, 344-350.

Singer, E. (2016). Reflections on surveys' past and future. *Journal of Survey Statistics and Methodology*, **4**, 463–475.

Sukhatme, P. V. (1953). *Sampling Theory of Surveys with Applications*. The Indian Society of Agricultural Statistics, New Delhi. Iowa State College Press, Ames, Iowa.

Sukhatme, P. V. and Sukhatme, B. V. (1970). *Sampling theory of surveys and its applications*. Iowa State College Press, Ames, Iowa.