STATISTICS IN TRANSITION new series, June 2023
Vol. 24, No. 3, pp. 1–22, https://doi.org/10.59170/stattrans-2023-029
Received – 08.05.2023; accepted – 31.05.2023

Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day

Graham Kalton[1]

Abstract

At the beginning of the 20th century, there was an active debate about random selection of units versus purposive selection of groups of units for survey samples. Neyman's (1934) paper tilted the balance strongly towards varieties of probability sampling combined with design-based inference, and most national statistical offices have adopted this method for their major surveys. However, nonprobability sampling has remained in widespread use in many areas of application, and over time there have been challenges to the Neyman paradigm. In recent years, the balance has tilted towards greater use of nonprobability sampling for several reasons, including: the growing imperfections and costs in applying probability sample designs; the emergence of the internet and other sources for obtaining survey data from very large samples at low cost and at high speed; and the current ability to apply advanced methods for calibrating nonprobability samples to conform to external population controls. This paper presents an overview of the history of the use of probability and nonprobability sampling from the birth of survey sampling at the time of A. N. Kiær (1895) to the present day.

Key words: Anders Kiær, Jerzy Neyman, representative sampling, quota sampling, hard-to-survey populations, model-dependent inference, internet surveys, big data, administrative records.

1. Introduction

This paper presents a selection of the major developments that have taken place over the years since social surveys were first introduced in the late 19th century. I restrict my coverage to surveys of households and persons and my focus is on the sampling methods used to conduct such surveys. Major changes have also taken place in modes of data collection, in questionnaire design, and in other aspects of survey research over the years, but these topics are outside the scope of this paper. My paper on the more general theme of survey research over the past 60 years overlaps with this paper and gives greater coverage on some topics (Kalton, 2019).

[1] Joint Program in Survey Methodology, University of Maryland, College Park, MD, USA.
E-mail: gkalton@gmail.com. ORCID: https://orcid.org/0000-0002-9685-2616.

The changes that have occurred in methods of survey sampling have arisen for many reasons, including developments in sampling theory, the continuing growth in computer power (that was non-existent for the first fifty years of survey research), new sampling frames, and the problems created by a broader and more challenging range of applications of social surveys that has occurred as the potential for survey research has been more fully recognized. While acknowledging these changes, it is noteworthy that many aspects of the sampling methods that have been superseded over time have remained relevant. Indeed, much of the current discussion of the use of nonprobability sampling and big data sources has roots in the early days of survey research.

Without attempting to date the origins of survey research, early applications of survey research for studying the social conditions of populations took off in the late 1800's. English examples include Charles Booth's large-scale survey of the social conditions of the population of London that was started in 1886, Seebohm Rowntree's survey of working-class poverty in York that was conducted a decade later, and Bowley's survey of working-class conditions in Reading in 1912, which he followed up with surveys in four other English towns (three of which were conducted by Burnett-Hurst under Bowley's direction). See Caradog Jones (1949) for the early surveys in England, Converse (2017) for an account of the history of survey research in the United States from its beginnings at the turn of the century through until 1960, and Stephan (1948) for a history of the use of sampling procedures dating back from earlier times through until the 1940's, primarily in the United States.

The London and York surveys were complete censuses of the surveys' target populations. As complete censuses, they were deemed statistically acceptable at the time; they were known as 'monographs' of their local communities. For the London survey, the target population was households with school-aged children, while for the York survey it was households that did not have servants (conducted only in streets that were likely to contain households without servants). Bowley had long argued for the use of sampling for such surveys, and he played a major role in its adoption (Aldrich, 2008). He used sampling for the first time in the five towns surveys, where systematic sampling was employed (Bowley, 1913), and he introduced the idea of measuring sampling errors for survey estimates.

As Kish (1995) notes, the emergence of the field of survey sampling can be dated from work led by the Norwegian statistician Anders Kiær, the first Director of Statistics Norway. Kiær developed a sampling method that he termed "representative sampling". Kiær's method of purposive sampling is worth reviewing both for the procedures he devised to make a sample nationally 'representative' and for the reactions to the method from statisticians attending meetings of the International Statistical Institute (ISI) at the time. The next section provides a brief overview of these issues.

2. Kiær's Representative Method of Statistical Surveys

Kiær's sampling methodology is described in detail in his monograph *The Representative Method of Statistical Surveys,* first published in Norwegian in 1897 and republished in 1976 with an English translation (Kiær, 1976). The monograph provides a good deal of detail on the sample designs Kiær developed for two large-scale surveys—one on personal income and property (PIP) and the other on living conditions (LC)—as well as reporting the objections to his methods that he received when he presented them at ISI meetings. As distinct from the surveys of English towns cited above, Kiær aimed to produce survey estimates for the whole of Norway. For this purpose, he developed two-stage area sample designs for his surveys: at the first stage, he selected a "representative" sample of administrative districts (rural districts or counties, towns, and cities); at the second stage, he drew samples of people for each survey. The choice of the sampled first-stage units was carefully fashioned to give geographical spread and to achieve a good representation of the Norwegian population in terms of characteristics collected in the 1891 Population Census (e.g., age, marital status, occupation, urbanicity).

The sample for the PIP survey was defined as men aged 17, 22, 27, etc. who had names starting with certain letters, selected from 1891 census records that were being processed at the time, with a total sample size of around 11,400 men. The sample size for the LC survey was around 80,000 adults. The sample size to be obtained in each selected rural county was specified based on calculations from census data; within selected counties, the enumerators were instructed to follow certain routes and to select different types of houses, but otherwise they were left to make the selections. In the smaller towns, every 9^{th}, 5^{th}, or 3^{rd} house was selected. An extra sampling stage was introduced in the largest towns. For example, the sample of houses in Oslo was selected within a sample of streets. Moreover, a higher proportion of the streets with larger populations was included in the sample, but this feature was counterbalanced by the selection of houses at a lower rate in the large streets.

The driving objective with Kiær's approach was to produce a representative sample that would constitute a microcosm of the Norwegian population. He invented some intricate methods to attempt to achieve this objective. His purposive selection of first stage administrative units sometimes incorporated ideas of probability proportional to size sampling and subsampling at different rates in compensation, thereby avoiding an excessive sample concentration in a few large districts. Similarly, his street sample in Oslo has the same feature. He also employed a simple 2:1 weighting adjustment to compensate for the smaller proportion of members of the rural population in the PIP survey. (Before the advent of computers, anything other than simple integer weighting adjustments would have been extremely difficult to routinely apply.)

Despite his thoughtful approach, Kiær encountered a great deal of criticism of his methods when he presented them to the ISI in 1895. The dominant criticism, however, was not of the representative method, *per se*, but rather of a sample-based enquiry rather than a complete enumeration. In the words of one strong critic, von Mayr: "We remain firm and say: no calculations when observations can be made". Kiær also made presentations on the representative method at the 1897, 1901, and 1903 ISI sessions, at which they were subjected to similar criticisms, together with another one. At the 1903 session, von Bortkiewicz reported the results of a significance test he had conducted that found that Kiær's representative samples were not truly representative. See Kruskal and Mosteller (1980) for a detailed account of the ISI sessions.

At the same time, Kiær expertise was under attack at home for the LC survey, which was conductedon behalf of a parliamentary labor commission to inform a very contentious social security act that would provide highly expensive disability insurance. A three-person "critique committee" was established to review the commission's major recommendation and its statistical basis. One committee member, the actuary Jens Hjorth, was extremely critical of Kiær's statistics, including the survey design, the representative sample design, and the analysis. The attacks on the statistics that Kiær's produced for the commission were forceful, extensive, and widely debated. In the end, based on the results of some new surveys, Kiær admitted that he had initially seriously underestimated the extent of disability. After that time, representative sampling for large-scale surveys disappeared in Norway. Lie (2002) provides an informative account of the rise and fall of Kiær's representative sampling method.

The ISI discussion of survey sampling fell into abeyance until 1924 when the ISI appointed a commission for studying the application of the representative method in statistics. By that time, the idea of a "partial investigation" was widely accepted. In its 1926 report (Jensen, 1926), the Commission concluded that a sample was acceptable if it was sufficiently representative of the whole. To satisfy this condition the sample could be produced either by random selection with equal probability or by purposive selection of groups with a representative overall sample. The report also recommended that the survey results should, wherever possible, be accompanied by an indication of the errors to which they are liable.

3. Neyman's Seminal Paper

In 1934, Neyman presented his classic paper comparing the methods of random and purposive selection to the Royal Statistical Society (Neyman, 1934). Covering more than the comparison, the paper contained a detailed discussion of a methodology for making inferences from random—or, more generally, probability—samples of finite populations, including providing a definition of a confidence interval in this context.

He also critically examined the assumptions made when using data from a purposive sample to produce an accurate estimate of a population parameter.

He discussed the sample design of purposive selection of groups used by Gini and Galvani in selecting a sample of records from the already-processed Italian General Census of 1921 that was to be used as the basis for later analysis. For their sample, Gini and Galvani (1929) selected a sample of twenty-nine of the 214 districts in Italy, balanced on seven covariables (note that departs from Kiær's stipulation that a large wide-spread sample of areas is needed). While the sample worked well for the averages of the control variables, it often failed to adequately represent the national population for other characteristics, and for the distributions of the control variables. These findings led them to raise questions about representative sampling.

Neyman's paper was a watershed for survey sampling, leading to widespread adoption of probability sampling, particularly by national statistical offices. It also led to the development of an extensive range of sampling methods and the associated theory applicable to a variety of practical survey problems, as described in the several texts on survey sampling that appeared in the 1950's. The many contributions of statisticians at the U.S. Census Bureau led by Morris Hansen are particularly noteworthy; see, for example, the two-volume text by Hansen, Hurwitz, and Madow (1953). Statisticians active in research on sample designs for agricultural surveys, such as Yates in England and Mahalanobis in India, also made important contributions to the advancement of the subject. The sampling text by Yates (1949) was among the first books on survey sampling methods. In 1950, Mahalanobis went on to establish and lead the famous socio-economic National Sample Survey (NSS) of India. An interesting feature of the NSS sample design was that the sample was composed of four replicate samples. The survey results were presented for each replicate separately as well as for the full sample, with the aim of communicating to readers an indication of the amount of sampling error in the survey estimates (see, for example, Mahalanobis, 1946). This was thus a forerunner of variance estimation using replication methods.

Note that perfect application of Neyman's design-based inference for probability sampling depends on:

- The availability of a sampling frame that provides complete coverage of the finite target population;
- A sample design that assigns known and non-zero selection probabilities to every element in the target population;
- Survey responses from every sampled unit; and
- The use of survey weights in the analysis to compensate for unequal selection probabilities.

Under these conditions (and assuming no response errors), survey estimates can be computed that are design-consistent estimates of the population parameters without the need to make any assumptions about the characteristics of the survey population. Model assumptions made about the population structure may be used to make the sample design more efficient or in the computation of the survey estimates, but the consistency of the survey estimates remains irrespective of the validity of the model. What the model assumptions do affect is the precision of the survey estimates. For example, in a stratified sample, if the sampling fraction in a stratum is set at a higher rate because the elements in a stratum are incorrectly modeled to be more variable, the (weighted) sample mean will still be unbiased, but it will be less precise than if the stratum element variance has been correctly modeled. Similarly, if a set of auxiliary variables $\boldsymbol{X}$ is available for all population elements, and a function of the x's, $f(\boldsymbol{X})$, is used as a working model to predict the survey variable y, then the finite population total may be estimated by

$$\hat{Y}_d = \Sigma_U \hat{f}(\boldsymbol{X}_i) + \Sigma_s w_i e_i, \tag{1}$$

where Σ_U and Σ_s denote summations over the population and sample respectively, $\hat{f}(\boldsymbol{X}_i)$ denotes the model estimate of y_i using the sample estimates of the unknown model parameters, $e_i = y_i - \hat{f}(\boldsymbol{X}_i)$, and the weight w_i is the inverse of element i's selection probability. By including the weighted estimate of the population total of the e_i's in this estimate, $\hat{Y}_d$ is a consistent estimator of the population total Y irrespective of the suitability of the working model; the choice of working model affects only the precision of the estimate $\hat{Y}_d$. This estimator is model-assisted, using the terminology coined by Särndal, Swensson, and Wretman (1992), but it is not model-dependent. For simple random sampling, Cochran (1953) gave an early example of a model-assisted estimator with the ratio estimator $\hat{Y} = (\bar{y}/\bar{x})X$, where X denotes the population total for the auxiliary variable x. An additional, important, feature of design-based inference is that estimates of the variances of sample estimates can be computed from the sample itself.

While the lack of dependence of design-based inference on model assumptions is the major attraction of probability sampling, it needs to be acknowledged that probability sampling is rarely perfectly executed in practice. There are two main sources of imperfection: noncoverage and nonresponse. Noncoverage, which arises because the sampling frame fails to include some elements of the target population, is widespread and its magnitude is often underrated. Area sampling is widely used in social surveys, selecting a probability sample of geographical areas, listing the households or dwelling units in the sampled areas, selecting a probability sample of households, and selecting either all or a probability sample of persons in those households. Even when the sample

of areas provides complete geographical coverage, noncoverage arises often from incomplete listing of households or dwelling units within sampled areas, and from incomplete listing of persons within sampled households. Nonresponse occurs when a sampled element fails to provide acceptable responses to some or all the survey questions. In the early years of probability sampling, response rates were high, and these two sources of imperfection were treated as minor blemishes that received little attention. They were either ignored or treated by simple weighting adjustments (simple, in part because more complex adjustments were computationally infeasible at the time).

Probability sampling has two main drawbacks to be balanced against the theoretical attractions of design-based inference: cost and timeliness. The extra costs of probability sampling include the costs of tracking down sampled individuals, including repeat calls when the individual is not initially available. When area sampling is used, the sampling costs also include the costs of listing units within sampled areas. For similar reasons, collecting survey data from a probability sample takes longer, making the production of the survey estimates less timely. Timeliness is important for all surveys, but particularly for surveys where the results are highly time-dependent, such as political polls, surveys of outbreaks of certain infections, and surveys of areas that have experienced a recent disaster.

A variety of less rigorous sampling methods are used in an attempt to apply a probability sampling approach to address these drawbacks. However, since all these methods require modeling assumptions, none of them can be classified as probability sampling. For convenience, they are called 'pseudo-probability' methods in what follows. In the early days of design-based inference, the quasi-probability sampling method known as quota sampling was widely used in market research and in other applications. That method is described in Section 4. Three other quasi-probability sampling methods are described briefly in Section 5.

4. Quota Sampling

To set the scene for the need for imposing quota controls on a sample of the general population, consider the infamous Literary Digest Poll of 1936. To forecast the outcome of the 1936 U.S. Presidential Election, the Literacy Digest mailed a questionnaire to a sample of ten million individuals selected from telephone directories, lists of automobile owners, and registered voters. The results obtained from the two million respondents indicated a clear-cut victory for Alf Landon with 57 percent of the vote, whereas in fact Franklin Roosevelt won with 61 percent of the vote. The upper-class bias of the sample, and of the respondents within the sample, is a major part of the explanation of the discrepancy between these percentages. No weighting adjustments

were employed to attempt to address the bias at the time. (Lohr and Brick, 2017, reweighted the sample using respondents' reports of their voting in the 1932 election, and these adjustments led to a correct prediction of the outcome, but the estimate of the vote for Roosevelt still fell far short of the actual vote.) This study serves to demonstrate that a large sample size does not necessarily yield good estimates. See Converse (2017) for more details.

Market researchers and pollsters developed the methods of quota sampling separately from the developments in probability sampling, with the aim of addressing the biases from uncontrolled sampling. There are various forms of quota sampling, with the essence of all of them being to control the types of persons to be interviewed. Interviewers are instructed to make their samples of respondents conform to specified quota controls by such characteristics as sex, age group, and employment status. The controls could be independent (e.g., so many men and so many women, so many persons over 35 and so many persons 35 years of age or less) or the numbers to be interviewed could be interrelated (e.g., so many men over 35, so many women over 35). Sudman (1966) describes a method of quota sampling for national face-to-face interview surveys that he termed "probability sampling with quotas". He employed the four quota control groups of men under 35, men 35 and older, employed women and unemployed women, with the control groups chosen to give appropriate representation to young men and employed women. See also Stephenson (1979). The interviewing field force would generally be distributed across the country in a balanced way, either in areas selected to be representative, along the lines employed by Kiær, or in areas selected by a probability sample design. Sometimes additional controls are imposed, for example specifying the routes the interviewers were to follow, with no more than one person sampled in any household. Quota controls can also be applied in telephone surveys, mall intercept surveys, internet surveys (see Section 6), and other types of survey.

Quota sampling has two main advantages over probability sampling: cost and timeliness. Quota sampling is less costly because interviewers do not need to chase up elusive sampled units and because it avoids the costs of sampling specific households or persons (often including the associated listing costs). For the same reasons, a quota sample can be speedily fielded, and the data collected more rapidly than with a probability sample.

Quota sampling is a form of nonprobability sampling that assumes that the respondents in a quota group are an equal probability sample of the population in that group. Note that this assumption also assumes that nonrespondents in the group are missing at random; nonresponse occurs with quota sampling, in essence with respondents substituted for the nonrespondents. Studies that have been conducted to evaluate quota sampling have found that the results are often similar to those produced

by probability sampling, but this is not always the case (see Moser and Stuart, 1953, also Moser and Kalton, 1971; Stephan and McCarthy, 1958). For further references on quota sampling, see Kruskal and Mosteller (1980).

Random Route Sampling. Random route, or random walk, sampling is another quasi-probability sampling method that avoids the cost of, and associated time involved with, the listing operation. There are various versions of this method, but each starts with a random selection of a starting household and the interviewers then follow specified rules for walking patterns to follow and selection methods to use for serially identifying the subsequent households. The method has often been used in Europe and it is used in the Expanded Programme of Immunization (EPI) sampling method described in Section 5. Bauer (2014, 2016) discusses the selection errors that can occur with random route sampling and demonstrates that the method does not produce an equal probability sample, as its users generally assume.

5. Pseudo-Probability Sample Designs for "Hard-to-Survey Populations"

Recent years have seen a major increase in the use of social survey methods to study the characteristics of "hard-to-survey populations" (Tourangeau, Edwards, Johnson, Wolter, Bates, 2014). Such populations are of various types, but all comprise only a small proportion of the general population and a population for which there is no separate sampling frame. This section presents three examples of sample designs for such populations. The first example is an inexpensive method that has been very widely used for vaccination surveys of the extremely rare population of 1-year-old children. The other two examples describe methods for sampling rare populations where membership of that population is a sensitive characteristic.

a. *The EPI sampling method.*

For almost 50 years, the World Health Organization's Expanded Programme on Immunization (EPI) has used simple, inexpensive, sample designs in developing countries for measuring childhood immunization at the district level. Many thousands of EPI surveys have been conducted over this period, and the sample design has evolved over time. The sample design is a two-stage sample of clusters of communities (e.g., villages, towns, health service districts) that are sampled with outdated measures of estimated population sizes, with samples of eligible children selected within selected communities. The standard overall sample size is small, with the selection of 30 clusters and 7 children in each cluster. The design is often known as 30 × 7 design. Except in smaller communities, no household listings are made. Instead, the interviewer goes to the center of the village, chooses a random direction by spinning a bottle on the ground, and counts the number of households in that direction to the edge of the

community. The interviewer then chooses a random number (for instance, from the numbers on a banknote) to identify the first sampled household. The second sampled household is then the one closest to the first, and so on, sequentially until survey data are collected on seven eligible children. Levy and Lemeshow (2008, pp. 427–428) describe the EPI sampling methods and Bennett (1993) describes some of the modifications to the original method.

The US Centers for Disease Control and Prevention (CDC) recommends a probability 30 × 7 sample design for its rapid needs assessment tool, the Community Assessment for Public Health Emergency Response (CASPER) program. In this case, the clusters are generally census blocks with counts of households obtained from the U.S. Census Bureau or by using a GIS program for use in the PPES selection of thirty clusters. The fieldworker counts or estimates the number of households in a sampled cluster, divides that number by seven to give the sampling interval for systematic sampling, proceeds to select the sample from a random starting point, selecting subsequent households using a serpentine walking procedure. A crude weighting adjustment is proposed for use in the data analysis. Details are provided by CDC (2019).

b. Venue-Based Sampling

Venue-based sampling (also known as location sampling, time-space sampling, center sampling, and intercept sampling) is used for sampling members of a rare population at places that they frequent. It is applicable for rare populations that visit certain locations. It can be used to survey nomadic populations and for sampling hidden rare populations where the membership of that population is a sensitive matter. The method requires the construction of a frame of locations and a decision on the overall time period for the survey, selecting a sample of location/time periods for data collection, and selecting all or a sample of members of the survey population visiting each sampled location in the sampled data collection time period (Kalton, 1991). Two issues of concern arise when sampling hidden populations. One relates to the population coverage provided by the frame of locations and the overall time period: What proportion of the population will fail to visit any of the locations in that time period? Another issue relates to the multiplicity problem: How to account for the variability in the numbers of visits made to any of the locations by different sample members during the overall time period? These numbers are needed for use in weighting to compensate for unequal selection probabilities, but they are unknown. At best, they can be estimated by asking respondents questions about their general frequencies of visiting the locations. See MacKellar, Gallagher, Findlayson, Lansky, and Sullivan (2007) for a description of the sampling methods used for surveying men who have sex with men (MSM) in a number of metropolitan areas in the United States.

c. Respondent Driven Sampling

Respondent driven sampling (RDS) is a form of link-trace sampling that selects the sample based on the social networks that exist for some populations. RDS has become a popular method for sampling rare hidden populations that have this feature, such as injection drug users and sex workers. The method starts by identifying a small set of members of the population of interest, who serve as *seeds* for the subsequent sample. The seeds respond to the survey, including responding to a question asking how many members of the survey population they know. They are then asked to recruit a set number of members of that population for the survey, the *alters*. The alters then go through the same process, recruiting further sample members. Under idealized circumstances, Heckathorn (1997) has shown that RDS produces a probability sample. However, the many conditions required for this to apply will not hold in practice (Gile and Hancock, 2010).

6. Internet Surveys

Recruiting the sample via the internet is a relatively recent approach for conducting social research. This approach has become extremely popular and has led to several alternative methods. See, for example, Baker, Blumberg, Brick *et al.* (2010) for a review of these methods. Surveys based on internet sampling have the great attractions of obtaining responses from large samples at low cost and high speed. However, their nonprobability sampling methods raise concerns about potential biases in the survey estimates. Those without, or with limited, access to the internet are excluded from these surveys and the survey respondents are clearly not a representative sample of the general population.

One form of internet sampling, known as river sampling, attaches invitations to participate in a survey on a number of internet sites, usually with offers of some form of compensation. The biases in the sample selection process make the representativeness of the sample highly questionable. Questions also need to be raised about the honesty and thoughtfulness of the responses.

Another form of internet sampling employs an opt-in internet panel. (An opt-in internet panel is distinct from an internet panel that selects a household panel by probability sampling and then conducts many data collections from the panel over time, albeit typically with low response rates). Extremely large numbers of people are recruited for opt-in internet panels to be available to be approached to respond to surveys over time, sometimes as one of a range of services they may be asked to provide, in exchange for a payment for their services. The panel members can then be selected for invitation to respond to a given survey based on their responses to the screening instrument used in their recruitment.

In some ways, these large-scale nonprobability internet surveys bring to mind the abysmal results obtained from the 1936 Literacy Digest Poll referred to early. However, there are two major differences from the uncontrolled sample in the Digest Poll. One is the attempt to select a representative quota sample in design with internet panels. The other is the use of weighting adjustments in the analysis to achieve the same purpose. Before around 1970, lacking today's computers, complex calibration weighting adjustments were infeasible, but now advanced adjustment methods have been developed and are readily employed for both probability samples (particularly those with low response rates) and for nonprobability samples. With river sampling, a limited number of variables can be collected as part of the data collection for use in calibrating the sample to known or estimated population characteristics. The data collected in the screening instrument for an on-line panel can provide a much greater range of variables that can be used in sample selection and in the application of complex calibration adjustments to make the weighted sample correspond to a wide range of external controls. Nevertheless, serious doubts will persist about whether external data are available for the key auxiliary calibration variables at the population level or for a probability sample of that population, and whether the responses to the on-line survey can be treated as equal to the responses from the external source. Thus, for any given survey estimate, there must be concerns about how representative the nonprobability sample members are of the general population within the controls imposed in design or weighting. There will inevitably remain some residual biases of unknown magnitude and, with large samples, these biases can have a dominant influence on the level of accuracy of the survey estimates (Meng, 2018; Kalton, 2021, pp. 136–137).

7. Model-Dependent Inference

In 1976, Fred Smith—my late friend and colleague at the University of Southampton at that time—wrote a paper reviewing the foundations of survey sampling in which he raised the question of why finite population inference should be so different from inference in the rest of statistics. His view at the time was that 'survey statisticians should accept their responsibility for providing stochastic models for finite populations in the same way as statisticians in the experimental sciences' (Smith, 1976); he moderated his position in a subsequent paper (Smith, 1994). Smith (1976) and papers by Brewer (1963), Royall (e.g., 1970, 1976) and others led to a spirited and longstanding debate about the choice between design-based (model-assisted) inference and model-dependent (or model-based) inference. I was a discussant of Fred's 1976 paper and I subsequently published two papers on the role of models in survey sampling inference, with a defense of design-based inference in most circumstances applicable in large-scale social surveys (Kalton, 1983, 2002). However, models are needed to deal

with the sampling imperfections of noncoverage and nonresponse, and they are needed for subgroup analyses in which the sample sizes are not adequate to provide design-based estimators of adequate precision. With the large decline in response rates that has occurred since the 1970's, it is no longer possible for survey statisticians to treat nonresponse as a minor blemish that can be brushed under the carpet in using design-based inference. I will return to this point later.

The model-dependent approach has led to the development of the prediction approach to survey inference. With this approach, an estimate of the population total Y is given by

$$\hat{Y}_m = \Sigma_{i\in s}^{n} y_i + \Sigma_{i\notin s}^{N} \hat{f}(\boldsymbol{X}_i) \tag{2}$$

where the first summation is over the observed values in the sample s of size n and the second summation is over the model predictions of the y values for the nonsampled elements in the population. For comparison with the model-assisted design-based estimator $\hat{Y}_d$ in (1), the model-dependent estimator may be expressed as $\hat{Y}_m = \Sigma_s e_i + \Sigma_U \hat{f}(\boldsymbol{X}_i)$. In practice, greater care is used to develop the model for $\hat{Y}_m$ than is the case in developing the working model for $\hat{Y}_d$. If the same model is used, $\hat{Y}_m$ likely has lower variance than $\hat{Y}_d$. However, $\hat{Y}_m$ has a design bias if the model is mis-specified, as is always the case to some extent, and the magnitude of the bias is unknown. The texts by Valliant, Dorfman, and Royall (2000) and Chambers and Clark (2012) describe the prediction approach in detail. The first chapter of Valliant et al. (2000) provides a useful review of design-based and model-based inference and includes further references. Note that the equation for $\hat{Y}_m$ does not include selection probabilities (except possibly for estimating the model parameters) and does not require a probability sample. However, as Valliant, Dorfman, and Royall (2000, pp. 19–22) argue, randomization has the benefit of giving some protection against imbalance in factors uncontrolled in the design.

In my experience, until recently the prediction approach has had limited utility for large-scale social surveys of households and persons for the following reasons:

1. As distinct from surveys of establishments, there are generally little, if any, data available from the sampling frame about every member of the target population for use in the prediction models. Although some countries maintain up-to-date population registers that contain a selection of individual characteristics, in many countries area sampling is used, with frame construction for individuals or households being performed only in selected areas. In these latter countries, no frame data is available for all members of the target population.
2. Social surveys are multipurpose in nature. They collect survey data on many variables, often numbering in the hundreds, and these data are analyzed in many ways, producing thousands of estimates. As a rule, these surveys are

primarily conducted to produce descriptive estimates of parameters of the survey's finite population. These estimates need to be produced rapidly and to be consistent with each other. (These days, analytic estimates are also often produced, mostly through secondary analyses–see section 7).

3. A large proportion of the variables collected in social surveys are categorical in nature. They often cannot be as well predicted from auxiliary data as is the case with some of the continuous variables collected in business surveys.

However, even with large-scale social surveys, model-dependent estimation has a role to play in the production of descriptive estimates for small subclasses for which the sample sizes are too small to yield design-based estimates of adequate precision. This situation occurs particularly when the subclasses are geographical-defined administrative areas. The growth of interest by policy makers and others in separate estimates for administrative districts of all sizes has led to the development of the subject known as *small area estimation*. For many years, small area estimates, which are obtained using model-dependent prediction methods, were viewed with considerable skepticism by design-based statisticians but they have now become widely accepted in many fields of application. Ghosh (2020) gives a history of the development of small area estimation over five decades and Rao and Molina (2015) give a detailed description of this large and growing field.

The theoretical developments in model-based inference have now become increasingly relevant for social surveys to address the sampling imperfections and limitations with probability samples, and for the analyses of nonprobability samples; the use of nonprobability sampling for social research has grown rapidly in recent years, in particular for internet surveys.

8. Analytic Uses of Survey Data

As computing power and software came into widespread use in the 1970's, survey data collected using complex sample designs were used, mostly in secondary analyses, to produce analytic statistics that studied the relationships between variables, often looking for causal connections. Initially, multiple regression was the main form of analysis, with interest directed to the magnitude of the regression coefficients. Many analysts argued that their interest in the results of these analyses was not for the specific finite population surveyed, but rather as estimates of superpopulation parameters of universal generality, and that, with the "correct" model, aspects of the sample design were irrelevant. From this perspective, probability sampling of the finite population becomes irrelevant and, unless survey weights and clustering were important as predictor variables, their inclusion in the analysis in a standard design-based way serves only to lower the precision of the estimated regression coefficients. The counter

position was that no model is totally correct and that the estimation of the population regression coefficients, often termed census parameters, using the survey weights provides a safer approach. There is extensive literature on this topic. See, for example, DuMouchel and Duncan (1983).

Over time, the use of regression methods with survey data has been extended to include a wide range of regression models and other multivariate analysis techniques such as categorical data analysis, multilevel modeling, and longitudinal analyses. It is outside the scope of this paper to describe the application of these methods with complex survey data. See Skinner, Holt, and Smith (1989), Chambers and Skinner (2003). Applications of a range of multivariate methods with complex survey data are well described in the texts by Korn and Graubard (1999) and Heeringa, West, and Berglund (2017).

9. Administrative Records and Big Data

A great deal of attention has been paid recently to the use of administrative records as an alternative source of research data. There are obvious serious issues of privacy and confidentiality to be addressed when government-maintained administrative data are used in this way. For this reason, this approach is particularly suited to researchers in government agencies. The approach has notable potential attractions in terms of cost and sample size, but it needs to be recognized that it has its limitations. For instance, what is the coverage of the frame of the records, especially regarding program enrollment versus eligibility? Do the records contain the data needed to measure the concepts as the researcher would like to define them? Are the record data measured consistently across the population, or are there differences in the procedures used in different administrative areas? Are the data measured consistently over time to enable time series data to be validly analyzed? How might changes in program rules affect temporal comparisons? How long is the period between data collection and the researcher's access to an analyzable dataset? Do the records contain the full set of variables needed for the analyses? In many cases, a single set of administrative records does not contain all the variables needed for the analyses. In this situation, it may be possible to link two or more sets of records, but record linkage problems need to be overcome and greater issues of confidentiality must be addressed.

How accurate are the data recorded in the records? Survey researchers have devoted a great deal of effort to training a relatively small number of interviewers to ask and record respondents' answers in a standard way. The situation is different with administrative records. Charlie Cannell, my late friend and colleague at the University of Michigan's Survey Research Center, had the following quotation from Josiah Stamp (1880–1941) in a plaque on his office wall:

> "The government are very keen on amassing statistics. They collect them, add them, raise them to the *n*th power, take the cube root and prepare wonderful diagrams. But you must never forget that every one of these figures comes in the first instance from the village watchman, who just puts down what he damn pleases."

While not claiming that current administrative records are as bad as this quotation might suggest, those who use such records for statistical purposes should carefully assess their quality and the distortions to which they may be subjected. See the paper by Hand (2018) and the ensuing discussion for a detailed discussion of the advantages and limitations of administrative records for research purposes.

In addition to government-maintained administrative records, there are other sources of social research data. In some cases, nongovernment records, such as those maintained by private organizations, may contain relevant information. However, they are subject to similar quality concerns, and access to the records may be hard to obtain. There are also sources of big data that occur on a flow basis, such as from linking cell phones to their GPS locations. The cell phone locations can be used to provide information about commuter times and even about long-distance travel trips if the home location is identified. Another source of big data is from scrapings on the web. Google Flu Trends (GFT) is a well-known and cautionary example. By analyzing extremely large numbers of flu-related searches on the web, Google developed models to predict local flu outbreaks in real time, avoiding the inevitable delay with other data sources. However, the models have since been found to fail (Lazer, Kennedy, King, and Vespignani, 2014), which serves as a warning that the apparent attraction of very big datasets can be illusory. For another example, see Bradley, Kuriwaki, Isakov, Sejdinovic, Meng, and Flaxman (2021).

10. Concluding Remarks

As illustrated in previous sections, the choice between purposive selection and probability sampling was a subject of debate in the early period of survey research. It was not until after Neyman's (1934) paper that probability sampling and design-based inference were established as the gold standard for large-scale surveys conducted by national statistical offices. With a perfectly executed probability sample and no response error, the analyst has the security of being able to report the survey findings as being subject only to a measurable degree of sampling error, whereas with nonprobability sampling the analyst can always be challenged that a purposive sample is not representative of the population with respect to the variables of analytic interest.

The preeminence of probability sampling for government surveys in the years from 1940 to, say, 2010 was not universal. There are costs incurred with probability sampling

and a probability sample takes more time to draw and data collection takes longer. As illustrated in earlier sections, failures to devise probability sampling methods that can be applied with acceptable cost and timeliness for certain populations has given rise to the development of shortcut methods that depart in varying degrees from rigorous probability sampling.

In the early days, the idea of a "representative sample" was restricted to a sample that was representative in its design, as was the case with Kiær's designs. The use of weighting adjustments in the analysis to achieve representativeness was seldom considered. The failure of the Literacy Digest poll in predicting the result of the U.S. Presidential election made clear that an extremely large unrepresentative sample could, without weighting adjustments, yield bad results.

Over the years, the implementation of probability sampling in social surveys has been increasingly challenged in many—but not all—countries by a steady decline in the willingness of the public to participate in surveys. Despite greater efforts to encourage response, response rates have declined dramatically in recent years. In reaction, greater efforts have been made to compensate for nonresponse, with major advances in the techniques employed. While replication methods of variance estimation can be applied to reflect the effect of the use of these techniques on the precision of the survey estimates, their use results in lower precision. Furthermore, the nonresponse adjustment model cannot be assumed to be "correct," and the extent of any remaining nonresponse bias cannot be assessed. With its current heavy reliance on nonresponse models, in many countries probability sampling with design-based inference no longer retains its status as the undisputed gold standard. Moreover, the current levels of nonresponse have led to a marked increase in the costs of conducting a survey with probability sampling, both because of the increase in the initial sample size needed to produce the required sample size and because of the increased efforts to counteract nonresponse. For example, in the U.S. random digit dialing (RDD) was widely used with telephone surveying in the later part of the last century and the early part of this one because of the cost-efficiency of this modality (particularly for surveying rare populations). However, response rates for RDD surveys have plummeted to a level as low as 10 to 20 percent, largely ruling out this form of sampling.

With the security of model-free probability sampling with design-based inference now a thing of the past, model-dependent methods appear to be taking on a major role in social statistics. Research on making valid inferences from nonprobability samples is ongoing (see, for example, Valliant, 2020). Models are increasingly used to analyze data from a combination of data sources, including survey data from probability and nonprobability samples, administrative records, and other sources of big data. Thus, there is much research currently underway on making inferences from combinations

of probability and nonprobability samples and from probability samples and other data sources (Kim and Wang, 2019; Beaumont and Rao, 2021; Rao, 2021),

In summary, after a long period in which probability sampling methods have dominated, the current situation is in a state of flux. New methods involving nonprobability sampling, internet sampling, administrative records, and big data are under constant modification and development. Brackstone (1999) lists six aspects of data quality for a statistical agency that remain applicable: relevance (how well the data meet the needs of the clients); accuracy (including both bias and variance); timeliness (time between the reference point and the time of data availability); interpretability (availability of relevant metadata); and coherence (ability to bring the data into a broader framework, including over time). The new data collection methods need to be assessed against these measures and, furthermore, the extensive research on response errors that has been conducted in the past now needs to be applied with the new methods of data collection. This is an exciting and challenging time for survey methodologists.

References

Aldrich, J., (2008). Professor A. L. Bowley's theory of the representative method. (Discussion Papers in Economics and Econometrics, 801) University of Southampton. https://eprints.soton.ac.uk/150493.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, G., Groves, R. M., Kennedy, C., Krosnick, J., Lavrakas, P. J., (2010). AAPOR Report on Online Panels. *Public Opinion Quarterly*, 74(4), pp. 711–781.

Bauer J. J., (2014). Selection errors of random route samples. *Sociological Methods and Research*, 43(3), pp. 519–544.

Bauer J. J., (2016). Biases in random route surveys. *Journal of Survey Statistics and Methodology*, 4(2), pp. 263–287.

Beaumont J-F., Rao, J. N. K., (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, 83, pp. 11–22.

Bennett, S., (1993). Cluster sampling to assess immunization: a critical appraisal. *Bulletin of the International Statistical Institute, 49th Session*, 55(2), pp. 21–35.

Bowley, A. L., (1913). Working-class households in Reading. *Journal of the Royal Statistical Society*, 76(7), pp. 672–701.

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., Flaxman, S., (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature,* 600, pp. 695–700.

Brewer, K. R. W., (1963). Ratio estimation in finite populations: some results deducible from the assumption of an underlying stochastic process, *Australian Journal of Statistics*, 5, pp. 93–105.

Caradog Jones, D., (1949). *Social Surveys.* Hutchinson's University Library, London.

CDC, (2019). Community Assessment for Public Health Emergency Response (CASPER) Toolkit. 3rd ed., CDC, Atlanta. https://www.cdc.gov/nceh/casper/.

Chambers, R., Clark, R., (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press, Oxford.

Chambers, R. L., Skinner, C. J., Eds., (2003). *Analysis of Survey Data.* Wiley, Chichester.

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

Converse, J. M., (2017). *Survey Research in the United States: Roots and Emergence 1890-1960.* Routledge, New York.

DuMouchel, W. H., Duncan, G. J., (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78, pp. 535–543.

Ghosh, M., (2020). Small area estimation: its evolution in five decades (with discussion). *Statistics in Transition*, 21(4), pp. 1–67.

Gile, K. J., Hancock, M. S., (2010). Respondent-driven sampling: an assessment of current methodology. *Sociological Methodology*, 40(1), pp. 285–327.

Gini, C., Galvani, L., (1929). Di una applicazione del metodo representative. *Annali di Statistica*, 6(4), pp. 1–107.

Hand, D. J., (2018). Statistical challenges of administrative and transaction data (with discussion). *Journal of the Royal Statistical Society, A,* 181(3), pp. 555–605.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample Survey Methods and Theory. Volume 1: Methods and Applications. Volume II: Theory.* Wiley, New York.

Heckathorn, D. D., (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems*, 44(2), pp. 174–199.

Heeringa, S. G., West, B. T., Berglund, P. A., (2017). *Applied Survey Data Analysis*. Chapman & Hall/ CRC, Boca Raton, FL.

Jensen, A., (1926) The report on the representative method in statistics. *Bulletin of the International Statistical Institute*, 22, pp. 355–376.

Kalton, G., (1983). Models in the practice of survey sampling. *International Statistical Review*, 51, pp. 175–188.

Kalton, G., (1991). Sampling flows of mobile human populations. *Survey Methodology*, 17(2), pp. 183–194.

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

Kalton, G., (2019). Developments in survey research over the past 60 years: A personal perspective. *International Statistical Review*, 87 (S1), pp. S10–S30.

Kalton, G., (2021). *Introduction to Survey Sampling*. 2nd ed. SAGE Publications, Thousand Oaks, California.

Kiær, A. N., (1976). *The Representative Method of Statistical Surveys*. English translation, Statistik Centralbyro, Oslo.

Kim, J. K., Wang, Z., (2019). Sampling techniques for big data analysis. *International Statistical Review*, 87 (S1), pp. S177–S191.

Kish, L., (1995). The hundred years' war of survey sampling. *Statistics in Transition*, 2(5), pp. 813–830.

Korn, E. L., Graubard, B. I., (1999). *Analysis of Health Surveys.* Wiley, New York.

Kruskal, W., Mosteller, F., (1980). Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review*, 48(2), pp. 169–195.

Lazer, D., Kennedy, R., King, G., Vespignani, A., (2014). The parable of Google flu: Traps in big data analysis. *Science*, 343, pp. 1203–1205.

Levy, P. S., Lemeshow, S., (2008). *Sampling of Populations. Methods and Applications.* 4th ed. Wiley, Hoboken, NJ.

Lie, E., (2002). The rise and fall of sampling surveys in Norway, 1875–1906. *Science in Context,* 15(3), pp. 385–409.

Lohr, S. L., Brick, J. M., (2017). Roosevelt predicted to win: Revisiting the 1936 Literary Digest Poll. *Statistics, Politics, and Policy,* 8(1), pp. 65–84.

MacKellar, D. A., Gallagher, K. M., Findlayson, T., Sanchez, T., Lansky, A., Sullivan, P. S., (2007). Surveillance of HIV risk and prevention behaviors of men who have sex with men—a national application of venue-based, time-space sampling. *Public Health Reports*, 122 (1), Supplement 1, pp. 39–47.

Mahanalobis, P. C., (1946). Recent experiments in statistical sampling in the Indian Statistical Institute (with discussion). *Journal of the Royal Statistical Society*, 109, pp. 325–378.

Meng, X-L., (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12(2), pp. 685–726.

Moser, C. A., Kalton, G., (1971). *Surveys Methods in Social Investigation*. 2nd ed. Heinemann, London.

Moser, C. A., Stuart, A., (1953). An experimental study of quota sampling. *Journal of the Royal Statistical Society, A,* 116, pp. 349–405.

Neyman, J., (1934). On 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, pp. 558–625.

Rao, J. N. K., (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B,* 83, pp. 242–272.

Rao, J. N. K., Molina, I., (2015). *Small Area Estimation*. 2nd ed. Wiley, Hoboken, N. J.

Royall, R. M., (1970). On finite population sampling theory under certain regression models. *Biometrika*, 57, pp. 377–387.

Royall, R. M., (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association,* 71, pp. 657–664.

Särndal, C.E., Swensson, B., Wretman, J., (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Skinner, C. J., Holt, D., Smith, T. M. F., Eds., (1989). *Analysis of Complex Surveys.* Wiley, Chichester.

Smith, T. M. F., (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society*, A, 139, pp. 183–204.

Smith, T. M. F., (1994). Sample surveys 1975-90; an age of reconciliation? *International Statistical Review*, 62, pp. 5–34.

Stephan, F. F., (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43(241), pp. 12–39.

Stephan, F. F., McCarthy P. J., (1958). *Sampling Opinions. An Analysis of Survey Procedures*. Wiley, New

Stephenson, C. B., (1979). Probability sampling with quotas: An experiment. *Public Opinion Quarterly*, 43(4), pp. 477–497.

Sudman, S., (1966). Probability sampling with quotas. *Journal of the American Statistical Association*, 61, pp. 749–771.

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K.M., Bates, N., Eds., (2014). *Hard-to-Survey Populations.* Cambridge University Press, Cambridge, U. K.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2), pp. 231–263.

Valliant, R., Dorfman, A. H., Royall, R. M., (2000). *Finite Population Sampling and Inference. A Prediction Approach.* Wiley, New York.

Yates, F., (1949). *Sampling Methods for Censuses and Surveys*. Griffen, London.