Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Three controversies in the history of survey sampling

by Ken Brewer

January 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2013
Vol. 39, No. 2, pp. 249-262
Statistics Canada, Catalogue No. 12-001-X

Three controversies in the history of survey sampling

Ken Brewer[1]

Abstract

The history of survey sampling, dating from the writings of A.N. Kiaer (1897), has been remarkably controversial. First Kiaer himself had to struggle to convince his contemporaries that survey sampling itself was a legitimate procedure. He spent several decades in the attempt, and was an old man before survey sampling became a reputable activity. The first person to provide both a theoretical justification of survey sampling (in 1906) and a practical demonstration of its feasibility (in a survey conducted in Reading which was published in 1912) was A.L. Bowley. In 1925, the ISI meeting in Rome adopted a resolution giving acceptance to the use of both randomization and purposive sampling. Bowley used both. However the next two decades saw a steady tendency for randomization to become mandatory. In 1934, Jerzy Neyman used the relatively recent failure of a large purposive survey to ensure that subsequent sample surveys would need to employ random sampling only. He found apt pupils in M.H. Hansen, W.N. Hurwitz and W.G. Madow, who together published a definitive sampling textbook in 1953. This went effectively unchallenged for nearly two decades. In the 1970s, however, R.M. Royall and his coauthors did challenge the use of random sampling inference, and advocated that of model-based sampling instead. That in turn gave rise to the third major controversy within little more than a century. The present author, however, with several others, believes that both design-based and model-based inference have a useful part to play.

Key Words: Rule of three; Representative method; *p*-statistic; Prediction; Randomization; Model, Horvitz-Thompson.

1 Introduction

One of the most difficult problems I struck in writing this paper was in knowing where to begin. Initially I had intended to start with Laplace as I had in an earlier paper (Brewer and Gregoire 2009), which incorrectly described him as being unable to fulfill his ambition to estimate the population of France by using what we would now describe as survey sampling. He had, in fact, achieved that using a sample of the small administrative districts known as communes as early as September 22, 1802 (Cochran 1978). In later accounts, I had read of him struggling to repeat that performance while the boundaries of France were in a constant state of flux, and I had jumped to the incorrect conclusion that he had never achieved it at all.

However, I soon found myself being pulled further back into history. No, Laplace had not been the first person to use a ratio estimator, not even the first Frenchman (Stephan 1948). The Englishman John Graunt had used the ratio estimator in his estimation of the population of London (Graunt 1662). Well, perhaps he had not really used the ratio estimator (he probably hadn't used anything that would be recognized as a ratio estimator today, certainly not by a finicky survey statistician like me!), but he had admittedly used the Rule of Three.

I had not come across that Rule before, but apparently it was well-known to be this: "If AB = CD and D is unknown, then D = AB/C." Obviously the present-day ratio estimator was a particular case of that Rule of Three. In fact, the Rule of Three must have predated the 17th Century by a considerable margin, so it might genuinely be of interest when searching for a survey-sampling start date.

1. Ken Brewer, School of Finance, Actuarial Studies and Applied Statistics, College of Business and Economics, Australian National University, Australia. E-mail: ken.brewer@anu.edu.au.

It soon occurred to me that the Rule of Three was bound to have been known to Hammurabi's astronomers, getting on for 4,000 years ago, because they were very arithmetically minded, having invented a sexagesimal system of counting that still survives today in the using of "hours" (and also of "degrees") "minutes" and "seconds", and also of 30^0-60^0-90^0 ["30-60-90"] triangles.

That realization encouraged me to start to look for a more recent starting point for this paper, and I eventually concluded that a good choice would be to start with "modern survey sampling", a topic that had been suggested to me before. The paper is structured as follows. Section 2 discusses the first controversy which is Anders Kiaer's "Representative Method." Section 3 provides a discussion on the second controversy, which is the exclusive use of randomization as a means for selecting samples, as advocated by Neyman (1934). The arguments for using the model-assisted or the model-based approach as a means for inference in survey sampling is described in Section 4. Section 5 provides a middle ground that incorporates both procedures. The paper ends with a summary given in Section 6.

2 The first controversy: Anders Kiaer and the "Representative method"

Anders Kiaer (1838-1919), was the founder and first director of Statistics Norway. Although many now claim him to be the first modern survey statistician, his contribution to statistics did not go unchallenged at the time. It was claimed, for instance, that his approaches to sampling lacked a theoretical description. In addition, there was also a serious lack of references in Kiaer's papers. Most of the charges made against him by his contemporaries have merit, but it is also true that with the first publication of his ideas in 1895 he started a process that ended in the development of modern survey sampling theory. Kiaer was also the first to use a sample survey on its own, as opposed to a by-product from a full enumeration.

By 1895, Kiaer had been conducting sample surveys successfully in his own country for fifteen years or more, finding to his own satisfaction that it was not always necessary to enumerate an entire population to obtain useful information about it. He decided that it was time to convince his peers of this fact, and he attempted to do so at the session of the International Statistical Institute (ISI) being held in Berne that year. Kaier there argued that what he called a "partial investigation", based on a subset of the population units, could indeed provide such information, provided only that the subset in question had been carefully chosen to reflect the whole of that population in miniature. He described this process as his "representative method", and he was able to gain some support for it, notably from his Scandinavian colleagues. Unfortunately, his idea of "representation" was too subjective and (in hindsight) too lacking in probabilistic rigour, to make headway against the then universally held belief that only complete enumerations, "censuses", could provide any useful information (Wright 2001, Lie 2002).

Moreover, all Kiaer's innovations, and in particular his idea of a sample being "representative", were controversial enough to create serious opposition to his ideas among his contemporaries, and this was particularly evident in the seriously unfavourable reactions to the paper that he presented at that 1895 meeting. However, he persisted and continued to present papers about his surveys and the methods he used in them at later ISI meetings.

Eight years later, at the ISI's Berlin meeting in 1903, Lucien March suggested that randomization might provide an objective basis for the use of "partial investigations" (Wright 2001, Lie 2002).

This idea was further developed by Sir Arthur Lyon Bowley, first in a theoretical paper (Bowley 1906) and later by a practical demonstration of its feasibility in a survey conducted in Reading, England (Bowley 1912).

By 1925, the ISI at its Rome meeting was sufficiently convinced (largely by the report of a study that it had itself commissioned!) to adopt a resolution giving acceptance to the idea of sampling. However it was left to the discretion of the investigators whether they should use randomized or purposive sampling. With the advantage of hindsight we may conjecture that, however vague their awareness of the fact, the writers of that report were intuiting that while purposive sampling was sometimes capable of presenting useful estimates, the underpinning of randomization was also desirable.

In the following year, Bowley himself published a substantial monograph (Bowley 1926) in which he presented what was then known concerning the purposive and randomizing approaches to sample selection, and also made suggestions for further developments in both of them. These included the notion of collecting similar units into groups called "strata," including the same proportions of units from each stratum in the sample. Furthermore, there was an attempt to make purposive sampling more rigorous by taking into account the correlations between the variables of interest for the survey and any other auxiliary variables that might be helpful in the estimation process.

3 The second controversy: Neyman advocates the exclusive use of randomization

By the 1920s the situation was clear, though hardly ideal. Sampling was no longer regarded as off the agenda, but there was little or no guidance as to whether the sample should be chosen randomly or purposefully. The next two decades saw a slow but steady tendency for the randomization approach to become mandatory. And there was a good reason behind that tendency, for there were no other attractive models available to cause sampling statisticians to want to use them.

A particularly influential paper advocating the exclusive use of randomization was Jerzy Neyman's (1934) 68-page attack on a survey conducted by Gini and Galvani (1929). Those two authors had selected a "purposive" sample of 29 out of 214 districts (circondari) from the 1921 Italian Population Census. Their sample was chosen in such a way as to reflect almost exactly the whole-of-Italy average values for seven variables chosen for their importance; but Neyman showed that it exhibited substantial differences for other important variables. He then went on to attack this study with a three-pronged argument.

1) Because randomization had not been used, the investigators had not been able to invoke the Central Limit Theorem. Consequently they had been unable to use the normality of the estimates to construct the "confidence intervals" that Neyman himself had recently invented. That idea appeared in English for the first time in this paper.
2) On Gini's and Galvani's own admissions, the difficulty of their achieving their "purposive" requirement (that the sample match the population closely on seven variables) had caused them to limit their attention to the 214 districts rather than to the 8,354 communes into which Italy had also been divided. In consequence, their 15% sample consisted of only 29 districts (instead of perhaps 1,200 or 1,300 communes). Neyman further showed that a considerably more

accurate set of estimates could have been expected had the sample consisted of a much larger number of those (order of magnitude smaller) communes.

3) Crucially, the population model used by the investigators was unrealistic and inappropriate. (Neyman was convinced that models by their very nature were always liable to represent the actual situation inadequately.) Furthermore, randomization obviated the need for such population modelling. Using randomization-based inference, the statistical properties of an estimator could be established by using the distribution of its estimates from all the samples that could possibly be drawn. Moreover, when using randomisation, that same estimator under different designs could have different statistical properties. (A good example of this, though not one of Neyman's, is that an estimator that is biased under an equal probability design might well be unbiased under an unequal probability design.)

These three arguments were not all equally valid or convincing, but even Gini and Galvani were ready to admit that something was seriously wrong with their approach. Moreover, the second argument (that the sample size of 29 was too small) was an easy one for Neyman to argue. It was incontrovertible. The third argument, that the population modelling was inadequate, was also one that the survey designers were ready to acknowledge. The first argument (about confidence intervals) seems to have been accepted for no better reason than that Neyman was saying it, and that since he was certainly right on the other two points, he was probably right on that one as well.

3.1 Bowley's opposition to Neyman's first argument and the outcomes

One statistician who was not prepared to accept Neyman's way of thinking was Bowley, who moved the vote of thanks to him for his 1934 presentation. We are, in consequence, able to quote the actual words used by both the disputants. Bowley actually started the argument by wondering aloud whether confidence intervals were just "a confidence trick"!

He asked, "Does [a confidence interval] really lead us to what we need—the chance that within the universe which we are sampling the proportion is within these certain limits? I think it does not. I think we are in the position of knowing that *either* an improbable event had occurred *or* that the proportion in the population is within these limits… The statement of the theory is not convincing, and until I am convinced I am doubtful of its validity."

In his reply, Neyman asserted that Bowley's question (about the confidence interval being a confidence trick) "contain[ed] the statement of the problem in the form of Bayes" and that in consequence its solution "*must* depend upon the probability law *a priori*." He added, "In so far as we keep to the old form of the problem, any further progress is impossible." He thus concluded that there was a need to stop asking Bowley's "Bayesian" question and instead adopt the stance that Neyman's own "*either…or*" statement [that *either* an improbable event had occurred *or* the proportion of the population was within the stated limits] "form[ed] a basis for the practical work of a statistician concerned with problems of estimation…"

However, the fact remains that confidence intervals are not easy to understand. A confidence interval is in fact a sample-specific range of potentially true values of the parameter being estimated, which has been constructed so as to have a particular property. This property is that, over a large number of sample observations, the proportion of times that the true parameter falls inside that range (constructed for each

sample separately) is equal to a predetermined value known as the confidence level. This confidence level is conventionally written as $p = 1 - \alpha$, where α is small compared with unity. Conventional values for α are 0.05, 0.01, and sometimes 0.001. Thus, if many samples of size n are drawn independently from a normal distribution, the proportion of times that the true parameter value will lie within any given sample's own confidence interval will, before that sample is selected, be $[1 - \alpha]$.

"It is not the case, however, that the probability of this true parameter value lying within the confidence interval as calculated for any individual sample of size n will be $[1 - \alpha]$. The confidence interval calculated for any individual sample of size n will, in general, be wider or narrower than average and might be centred well away from the true parameter value, especially if n is small. It is also sometimes possible to recognise when a sample is atypical and, hence, make the informed guess that in this particular case, the probability of the true value lying in a particular 95% confidence interval differs substantially from 0.95."

Let us then consider, in particular, the most commonly used of all 95% confidence intervals, namely that between $p = 0.05$ and $p = 1.00$. (Fisher (1925) had actually suggested using the interval between $p = 1/22$ and $p = 1$.) Editors of publications in a great variety of fields (most of them not themselves statisticians) feel this definition of "significance" to be the one that very conveniently gives them leave to publish *p*-values that fall outside that range and reject those that do not. I believe the time is long overdue for looking at that suggestion of Fisher's very carefully.

What Fisher claimed (using $p = 1/22$ rather than $p = 0.05$) was that "Using this criterion we should be led to follow up a false indication only once in 22 trials". But what did he (and what do we now) mean by "following up a false indication"? What we *should* mean is this: that if the null hypothesis (H_0) is true, a "false indication", that is to say, "a misleadingly significant observation," will be observed , on average, once in 22 (or 20) times. But this is not what many non-statistical users of the *p*-statistic imagine that it means. Such users seem to think it means that only one in 20 of their "significant observations" (*i.e.*, that only one in 20 of all their observations with *p*-values less than 0.05) will be *misleadingly* significant.

That is the notorious *p*-statistic fallacy! (See Berger and Sellke (1987) for details.) To say "If H_0 is true, observations will be misleadingly described as 'significant' only once in 20 (or 22) times", is correct but unhelpful, for if H_0 is true, it follows that *every* observation described as "significant", for whatever reason, must also have been described that way misleadingly. But simply to say "Whether H_0 is true or not, $p < 0.05$", is also misleading. A meaningful false discovery rate (FDR) in these circumstances is (in fact) something that approximates to $p < 0.0025$ or $p < 0.05^2$.

This is a subject on which I have expended some thought of late. In particular, I co-authored a four-part article on it.

Part 1 (Brewer and Hayes 2011a) discusses how the notoriously parsimonious Bayesian Information Criterion (BIC) can be remedied by adding certain obviously needed penalty terms. The resulting Augmented Bayesian Information Criterion (ABIC) is nearly always intermediate between the original BIC and the (equally notoriously *lacking* in parsimony) Akaike Information Criterion (AIC). Another useful feature of the ABIC is that in its univariate case it is a simple function of T (the large sample limiting case of Student's t).

In Part 2 (Brewer and Hayes 2011b), a reference Bayesian hypothesis test is derived that is fully compatible with the ABIC of Part 1. An important role is played here by an obvious generalisation of Benford's (purely empirical) Law of Numbers, in providing an objective (though not flat) Bayesian prior distribution over the entire range from zero (or minus infinity) to plus infinity for the relevant hypothesis test. (The problem that characteristically arises with zero prior probabilities is avoided here by the use of Lebesgue-type measures instead.) Importantly, when $T = 1$, the relevant Bayesian hypothesis test yields a posterior measure that is indifferent between the null and alternative hypotheses. Furthermore, when the ABIC is generalised to small samples, as a function of the *t*-statistic, Fisher's p sets an upper bound to the false discovery rate (FDR), regardless of the number of degrees of freedom involved.

In Part 3 (Brewer, Hayes, and Gillison 2012), a set of some 1,300 regression slopes from a biodiversity sample survey of tropical landscape mosaics is used to provide empirical support for the ABIC, and the earlier theoretical findings are thereby confirmed.

In Part 4 (Hayes and Brewer 2012), the approximate results derived in Parts 1 to 3 are supplemented by exact results that can be obtained using a somewhat similar approach, but one that requires no explicit null hypothesis. Finally we suggest some likely consequences of the recognition that, when the implied null hypothesis is precise, much smaller values of $|p|$ (typically of the order of 0.0025 rather than 0.05) are needed to provide any useful FDR.

3.2 The acceptance of Neyman's second and third arguments

The second and third ideas that Neyman had advocated in his paper (namely the inefficiency of Gini and Galvani's (1929) selection procedure and the need to use only randomized sampling) though both relevant for their time and well presented, caught on only gradually over the course of the next decade. W. Edwards Deming heard Neyman in London in 1936. He was impressed and arranged for Neyman to lecture, and for his approach to be taught to U.S. government statisticians. A crucial event in its acceptance was the use in the 1940 U.S. Population and Housing Census of a one-in-twenty sample, designed by Deming along with Morris Hansen and others, to obtain answers to additional questions. Once fully accepted, however, Neyman's second and third arguments swept all other considerations aside for at least two decades.

Those twenty-odd years were a time of great progress. In the terms introduced by Kuhn (1962), finite population sampling had found a universally accepted "paradigm" in randomization-based inference, and an unusually long period of "normal science" based on "probability sampling" had ensued. ("Probability sampling" requires that all the elements in the population have known and positive probabilities of inclusion in sample.)

3.3 The appearance of relevant textbooks

This agreed consensus made it possible for several influential sampling textbooks to be published. Kish's (1995) historical article mentions five that appeared in quick succession: Yates (1949), Deming (1950), Cochran (1953), Hansen, Hurwitz and Madow ("HH&M") (1953) and Sukhatme (1954).

In my estimation the two most important of these were those by Cochran and by HH&M, but for quite opposite reasons. HH&M seem not to have wanted any truck at all with population modelling. (I doubt

whether the word "model" is even mentioned in either of their two volumes. It does not appear in either index.) Cochran (1953), on the other hand found several uses for such models, even as early as 1953.

Re-reading Cochran (1953) recently, I had the distinct impression that the more he wrote, the more he was at ease in using population models. So I started to count them. This first edition had 316 pages of text. The words "model" and "models" were used on 23 occasions. In the first half of the book, the word "model" appeared only once (on page 123) and "models" not at all. But Cochran used those words again three times in the third quarter and 19 times in the last quarter. (Numbers sometimes speak louder than words!)

Another strange thing was that although HH&M's two-volume book on *Sample Survey Methods and Theory* appears not to have used the word "model" at all, each of its two volumes included a chapter on "regression estimation". I don't see how one can have a regression estimator without a regression model, at least in the back of one's mind.

HH&M also defined four "estimates" in Chapter 11 of their Volume 1: the *difference estimate*, the *regression estimate*, the *ratio estimate* and the *simple unbiased estimate*. In Chapter 11 of Volume 2 only the *difference estimate* and the *regression estimate* are defined, but of course the other two would have been well known to anyone who was already familiar with Volume 1.

The question still remains as to whether HH&M would have regarded the regression estimate as implying a model. My guess is that they would have been reluctant to do so!

3.4 My fifteen months in the USA

In 1966-67, I was privileged to spend over a year in the USA, visiting (in order) the U.S. Bureau of the Census in Washington DC, and then Harvard and Princeton Universities. At the Bureau of the Census I had hoped to be able to spend some time with Morris Hansen, and was looking forward to suggesting to him that there were actually some useful things that could be done with population models, but when the first opportunity occurred, he cut me off short, saying "We don't need *models*," and immediately changed the subject!

Conversely, when I went to Harvard, where I spent a considerable time with Cochran, we were able to look at the topic rationally together and agree that models had a useful if limited role to play. At Princeton, I attempted to interest several well-known statisticians at the university about the topic, but without any serious success.

Quite a different challenge to Hansen's model-free orthodoxy had been voiced by Godambe (1955), with his proof of the non-existence of any uniformly best randomization-based estimator of the population mean. A new notation and class of estimators were required for the argument, and this framework in its earliest form met with some resistance. In Section 5 of that paper, citing Yates' (1949) textbook and Cochran's (1939) paper as antecedents, Godambe suggested an alternative optimality criterion, the minimization of the expected sampling variance under what was later called a superpopulation model.

At that time few others working in this excitingly innovative field of survey sampling seemed to be concerned by this result. I must confess that I wasn't myself concerned at the time, but I now think that perhaps I should have been!

4 The third controversy: "Sampling inference: Model-assisted or model-based?"

It came as a considerable shock to the finite population sampling establishment when Royall (1970) issued his highly readable call to arms for the reinstatement of purposive sampling and prediction-based inference. To read this paper was to read Neyman (1934) being stood on its head. The identical issues were being considered but the opposite conclusions were being drawn.

By 1973, however, Royall had withdrawn the most extreme of his recommendations. This was that the best sample to select would be the one that was optimal in terms of a model represented by the following Equations:

$$Y_i = \beta X_i + U_i \tag{4.1}$$

$$E(U_i) = 0 \tag{4.2}$$

$$E(U_i^2) = \sigma^2 X_i \tag{4.3}$$

and

$$E(U_i U_j) = 0. \tag{4.4}$$

Such a sample would typically have consisted of the n largest units in the population as measured by their realized x_i values, asking for trouble if the parameter β had not been close to constant over the entire range of the sizes of the population units.

In later articles (Royal and Herson 1973a, Royal and Herson 1973b, Cumberland and Royall 1981), Royall suggested that the chosen sample be "balanced," in other words, that the moments of the sample x_i should be as close as possible to the corresponding moments of the whole population. This formalized the much earlier notion that samples should be chosen purposively to resemble the population in miniature. The samples of Gini and Galvani had been chosen in something of the same way – meaning here "something of the same way in intention", but certainly not anything like the same success in execution.

For the most part, Royall's original stand remained unshaken. The business of a sampling statistician was to make a realistic model of the relevant population, design a sample to estimate its parameters, and make all inferences regarding that population in terms of those parameter estimates. The randomization-based concept of defining the variance of an estimator in terms of the variability of its estimates over all possible samples was to be discarded in favour of the prediction-based variance, which was sample-specific, and based on averaging all possible realizations of the chosen prediction model.

Regardless of what sample was drawn, Royall's estimator for a population total $T_y = \sum_U y_i$ had this prediction form:

$$t_y = \sum_s y_i + \sum_{U-s} x_i \hat{\beta}_{\text{BLUE}},$$

where $\hat{\beta}_{\text{BLUE}} = \sum_s y_i \Big/ \sum_s x_i$ was the best linear unbiased estimator for β based on the sample under model in equation (4.1). This is in prediction form since the *y*-values of $U - s$ are predicted by the model.

Sampling statisticians had at no stage been slow to take sides in this debate. Now the battle-lines were drawn. The heat of the argument appears to have been exacerbated by language-blocks; for instance the words "expectation" and "variance" carried one set of connotations for randomization-based inference and quite a different set for prediction-based inference. So assertions made on one side appeared to those on the other side to be unintelligible nonsense.

A major establishment counter-attack was launched with an article by Hansen, Madow and Tepping (1983). A small (and by most standards undetectable) divergence from Royall's model was shown nevertheless to be capable of distorting the sample inferences substantially. The obvious counter would have been "But this distortion would not have occurred if the sample had been drawn in a balanced fashion."

5 A third alternative, "Use them both together"

Eventually, a third position was also offered, the one held by the present author, namely that since there were merits in both the design-based (or randomization-based) and the model-based (or prediction-based) approaches, and that since it was possible to combine them, the two should be used together. I had actually foreshadowed this possibility in Brewer (1963), a paper that provoked little interest at the time, but was later spotted and accorded recognition by J.N.K. Rao, at least to the extent that he invited me to visit him in Ottawa for six weeks in 1974.

To combine these two approaches was relatively simple. In each of them there was a variable y which was of central interest and a related or auxiliary variable x, about which something additional was known that could be of assistance in estimating the value of that y variable. That "something additional" was typically the known population total of all the x values, denoted by T_x. Consequently the *relationship* of central interest, was that which linked the crucial parameter β in equation (4.1) to its *cosmetic* estimator $\hat{\beta}_{\text{COS}}$, namely

$$\hat{\beta}_{\text{COS}} = \frac{\sum_s (\pi_i^{-1} - 1) y_i}{\sum_s (\pi_i^{-1} - 1) x_i}, \tag{5.1}$$

where π_i is the probability that unit i is selected in the sample, or in the notation used by Särndal (2011),

$$\hat{\beta}_{\text{COS}} = \frac{\sum_s (d_k - 1) y_i}{\sum_s (d_k - 1) x_i}, \tag{5.2}$$

where his d_k is identical to my π_i^{-1}. The resulting estimator of the total $Y = \sum_U y_k$ is

$$\hat{Y}_{\mathrm{COS}} = \sum_s d_k y_k + \left(\sum_U x_k - \sum_s d_k x_k \right) \frac{\sum_s (d_k - 1) y_k}{\sum_s (d_k - 1) x_k}. \qquad (5.3)$$

Särndal (2011) also shows that these x and y values can be related to each other in several different ways, but also shows that there is a common theme that runs through all of those ways. That common theme is that y increases linearly as x increases, and that the extent of that linearity is measured by the parameter β in equation (4.1). Importantly, however, when $\hat{\beta}_{\mathrm{COS}}$ replaces $\hat{\beta}_{\mathrm{BLUE}}$ in Royall's prediction estimator, the estimator can be shown to be nearly unbiased under the design regardless of the validity of the assumed model.

Equation (5.2) can also be found explicitly on page 569 of Brewer (2011), immediately following its more general formula in matrix notation, namely

$$\hat{\beta}_{\mathrm{COS}} = \left[X_s' Z_s^{-1} \left(\Pi_s^{-1} - I_n \right) X_s \right]^{-1} X_s' Z_s^{-1} \left(\Pi_s^{-1} - I_n \right) y_s . \qquad (5.4)$$

When, the question arises as to how many explanatory variables should be used in the relevant model, Särndal (2011) makes an apparently disparaging distinction between "explanatory rich" and "explanatory poor" countries. He certainly treats those "explanatory poor" countries as being at a substantial disadvantage as a result of having relatively few "explanators".

There is at least one "explanatory rich" country (Australia) that appears to have made a deliberate decision to ignore whatever advantages might be available to those that are "explanatory rich". The current Australian procedure (the one used primarily to produce seasonally adjusted series) is to use only a single auxiliary variable, namely the latest available Census total, as the single "explanator".

Earlier, Brewer (1999a) had also presented a case that it might be preferable to use a cosmetic regression estimator to compensate for any lack of balance, rather than go to the trouble of selecting balanced samples. However, those who prefer to use balanced sampling directly can now select randomly from among many balanced or nearly balanced samples using the "cube method" (Deville and Tillé 2004). That paper also contains several references to earlier methods of selecting balanced samples, but regardless of how the relevant balanced sample is arrived at, the ways in which it needs to be used are identical.

In Brewer and Gregoire (2009) all three of the relevant approaches to estimation (randomization alone, prediction alone, and the two together) are examined. At this point, it is convenient to quote from yet another paper of mine (Brewer 2005, pages 390-391) which sets out the reasons why I was, and still am, concerned to use both methods simultaneously, and how readily it can be done.

"Each approach has its merits, and there are advantages in using both together. Consider how each of these inferences works.

First, design-based inference. Consider the general case where the inclusion probabilities π_i are known but may differ from unit to unit. In that case we can imagine the sampling statistician constructing a model of the population by looking at each of the sample units in turn and saying, *Oh yes, you (the first unit) were included with one chance in 10, so my model of the population includes you and nine other non-sample units with the same Y_k value as you. But you (the second unit) you were included with only one chance in two, so my model includes you and only one other unit like you.*"

The consequence of using this procedure here was therefore that the model of the population in the sampler's mind would consist of two real sample units (one from each sample stratum) plus ten imaginary units, (nine from the stratum with a sample fraction of one in ten, plus one from the stratum with a sample fraction of one in two) and finally plus all the units from the completely enumerated stratum.

Brewer (2005, page 391) continues as follows: "So even design-based estimation can be thought of as being based on a model, but on a model quite different from the prediction models… that are favoured by the so-called *model-based* school. More accurately that school should be described as *prediction-based* and the *design-based* school should be described as *randomization-based*. Each school uses a model, but one uses a prediction model and the other a randomization model."

The randomization-based approach described above is the one that was used for the selection of two sample units (one from each sampled stratum) plus all the units in the completely enumerated stratum. It also gave rise to the well-known Horvitz-Thompson estimator, which may be written

$$\hat{T}_{\mathrm{HT}} = \sum_{i \in s} \frac{Y_i}{\pi_i} = \sum_{i=1}^{N} \delta_i \frac{Y_i}{\pi_i} \tag{5.5}$$

where δ_i is an inclusion indicator taking the value "one" if the i^{th} unit is either in the sample or in the completely enumerated sector, and the value "zero" otherwise. In this particular case it is defined over both the two sampled units and also all the units in the completely enumerated sector. [This last sentence corrects the error mentioned above.]

Statisticians of the prediction-based school ridicule the use of randomization-based inference because the inclusion probabilities are chosen arbitrarily by the sample designer, and are therefore unable (they say) to tell us anything meaningful about the population! They prefer instead to use the Best Linear Unbiased Estimator (BLUE) of the regression parameter β as a step towards arriving at the Best Linear Unbiased Predictor (BLUP) of T. It is a predictor, because T is a random variable under the model, not a parameter.

Which is then the better estimator of T, the HT or the BLUP? The BLUP is the better if the prediction model holds exactly, and is much the better if both the sample and the population are small. However there will always be some sample size beyond which the HT is the more efficient estimator unless the model holds exactly.

6 Summary

In conclusion, we can see that survey sampling, over its relatively short history, has been remarkably vulnerable to controversies. In the first instance there was opposition to the notion that there should be any sampling at all. The only valid source of statistical information was taken to be the complete collection. It took the determination of Kiaer, a person already in a senior position of authority, to break down the opposition to what was eventually demonstrated to be a valuable tool.

The second controversy was also due to the determination of just a few people. Neyman took the lead, but this time there were others who were involved. Bowley was certainly involved to start with, but Neyman seems to have had the more convincing arguments at the crucial time. They were controversial,

even to begin with, and I am certainly not impressed with them now, but at the time he found a ready disciple in Hansen, who dominated the sampling fraternity for decades, at least until the mid-1970s.

The third controversy is still in progress and it is not altogether clear as to how it will turn out, but my current preference (at least for middling-sized samples) would be to use the prediction and randomization estimators combined.

In summary, both the HT and the BLUP can be useful in different situations. The BLUP makes sense to use when the sample size is small, and a model is desperately needed. The HT provides protection against prediction-model failure as the sample grows large. A prudent statistician would combine the principles of both.

References

Berger, J.O., and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p*-values and evidence (with discussion). *Journal of the American Statistical Association*, 82, 112-139.

Bowley, A.L. (1906). Address to the economic and statistics section of the British association for the advancement of science. *Journal of the Royal Statistical Society*, 76, 672-701.

Bowley, A.L. (1912). Working class households in reading. *Journal of the Royal Statistical Society*, 76, 672-701.

Bowley, A.L. (1926). Measurement of the precision obtained in sampling. *Bulletin of the International Statistical Institute*, 22, 11-62 (supplement).

Brewer, K.R.W. (2011). Remarks on the paper on "Combined inference in survey sampling" by Carl-Erik Särndal. *Pakistan Journal of Statistics*, 27, 4, 567-572.

Brewer, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 10, 213-233.

Brewer, K.R.W. (1999a). Design-based or model-based inference? Stratified random *vs* stratified balanced sampling. *International Statistical Review*, 67, 35-47.

Brewer, K.R.W. (1999b). Cosmetic calibration with unequal probability sampling. *Survey Methodology*, 25, 205-212.

Brewer, K.R.W. (2005). Anomalies, probings, insights: Ken Foreman's role in the sampling inference controversy of the late 20th century. *Australian and New Zealand Journal of Statistics*, 47, 4, 385-399.

Brewer, K.R.W., and Gregoire, T.G. (2009). Introduction to survey sampling. Chapter 1 of *Handbook of Statistics* 29A, *Sample Surveys: Design, Methods and Applications*, (Eds., D. Pfefferman and C.R. Rao), Elsevier.

Brewer, K.R.W., and Hayes, G. (2011a). Understanding and using Fisher's *p*: Part 1: Countering the *p*-statistic Fallacy. *Mathematical Scientist*, 36, 107-116.

Brewer, K.R.W., and Hayes, G. (2011b). Understanding and using Fisher's *p*: Part 2: A Reference Bayesian Hypothesis Test. *Mathematical Scientist*, 36, 117-125.

Brewer, K.R.W., Hayes, G. and Gillison, A.N. (2012). Understanding and using Fisher's *p*: Part 3: Examining an Empirical Data Set. *Mathematical Scientist*, 37, 20-26.

Cochran, W.G. (1953). *Sampling Techniques.* First Edition, Wiley.

Cochran, W.G. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492- 510.

Cochran, W.G. (1978). Laplace's ratio estimator. In *Contributions to Survey Sampling and Applied Statistics*; papers in honor of H.O. Hartley; H.A. David (Editor), 3-10.

Deming, W.E. (1950). *Some theory of sampling.* Dover books on mathematics.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling, the cube method. *Biometrika*, 91, 893-912.

Fisher, R.A. (1925). *Statistical methods for research workers.* 14th Edition (1970) Oliver and Boyd.

Gini, C., and Galvani, L. (1929). Di una applicazione del metodo rappresentativo all' ultimo censimento italiano della popolazione (1 dicembre 1921). *Annali di statistica* VI 4, 1-107.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17, 269-278.

Graunt, J. (1661/2). Natural and political observations made upon the Bills of Mortality. Reprinted (1939) Baltimore: The John Hopkins Press.

Hansen, M.H., Hurwitz W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory (2 vols.)* (Republished 1993) Wiley, New York.

Hansen, M.H., Madow W.G. and Tepping, B.J. (1983). An evaluation of dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Hayes, G., and Brewer, K.R.W. (2012). Understanding and using Fisher's *p*: Part 4: Do we even need to specify a prior measure at H_0? *Mathematical Scientist*, 37, 27-33. Sons, New York.

Kiaer, A.N. (1897). The representative method of statistical surveys. Papers from the Norwegian Academy of Science and Letters, II The Historical, philosophical Section, 1897 No. 4.

Kish, L. (2003). Selected Papers. Graham Kalton (Editor) Steven Heeringa (Editor) Wiley.

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

Kuhn, T.S. (1962). The Structure of Scientific Revolutions. University of Chicago Press.

Lie, E. (2002). The rise and fall of sample surveys in Norway, 1875-1906. *Science in Context*, 15 (3), 385-1906.

Neyman, J. (1934). On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

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

Royall, R.M., and Herson, J. (1973a). Robust estimation in finite population I. *Journal of the American Statistical Association*, 68, 880-889.

Royall, R.M., and Herson, J. (1973b). Robust estimation in finite population II: Stratification on a size variable. *Journal of the American Statistical Association*, 68, 890-893.

Särndal, C.-E. (2011). Combined inference in survey sampling. *Pakistan Journal of Statistics*, 27 (4) 359-370.

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

Sukhatme, P.V. (1954). Sampling theory of surveys: With applications. Asia Publishing House.

Wright, T. (2001). Selected moments in the development of probability sampling: Theory and practice. *Survey research methods section newsletter*, American Statistical Association, Alexandria, VA. Issue 13, 1-6.

Yates, F. (1949). Sampling Methods for Censuses and Surveys, London, C. Griffin.