Why are design in survey sampling and design of randomised experiments separate areas of statistical science?

Dan Hedlin

1

1 Introduction

It is puzzling that design of randomised experiments and sampling design have diverged and become two separate areas of statistical science with weak interaction. As Kish (1965, p. 595) put it: “the separation of the sample surveys from experimental designs is an accident in the recent history of science...”. Lack of communication is certainly a factor. However, the origins of randomised experiments and survey sampling are different.

1Department of Statistics, Stockholm University
We identify two roots of survey sampling and we may call them ‘mathematical’ and ‘arithmetical’ (Bowley, 1906). One strand of survey sampling stems from population studies with one leg in arithmetical work limited to tabulation of population records (e.g. Swedish 18th century population statistics) and another leg in more ‘mathematical’ work in which sources of uncertainty are accounted for. There were also researchers who very early began to use random sampling. For example, Laplace invented and constructed a survey sampling method to estimate the population in France in the beginning of the 19th century. He used the central limit theorem to assess the accuracy of his estimates. His work on survey sampling was long buried in oblivion (Bethlehem, 2009). More impact did the work of Bowley (1913) have. He describes a sample survey of working-class households in Reading. Through a systematic sample (not clear if it is a random systematic sample or not) he estimates average household income and other population parameters. Other strands of early survey sampling have their roots in crop yield estimates and in opinion polls (Stephan, 1948).

To say that experiments also have two roots, one ‘refined common sense’ and one ‘mathematical’, may be to draw the parallel to surveys too far. However, experiments have a long and widely spread history. Sven Ove Hansson (2007, p.48) notes that “the Mende people in Sierra Leone have a special word, hungoo, for experiment. A hungoo can consist in planting two seeds in adjacent rows, and then measuring the output in order to determine which seed was best. This was probably an original habit, not one brought to the Mende by visiting Europeans. Similar experiments also occur in other parts of the world.” Those experiments may well be denoted ‘refined common sense’, whereas Fisher (1935) certainly treated experiments ‘mathematically’. But the tenuousness in the parallel to surveys may sit in the word ‘roots’: Fisher elaborated on existing practices of experiments.

In this article I attempt to cover large areas of statistical practice and history. Since I cannot possibly be deeply familiar with all aspects of the history of survey sampling and randomised experiments and the contemporary methodology, I will appreciate suggestions for correction and addition.
2 What is an experiment?

In Sven Ove Hansson’s (2007) account an experiment, generally speaking, consists of four steps:

1. Make it happen, that is, make the phenomenon that you intend to study appear.
2. Separate. Disturbing factors needs to be removed.
3. Control. Factors that may influence the phenomenon must be controlled.
4. Observe. Obtain as exact observations as possible.

A fifth step would be to select the objects that will be included in the study. This fifth step could be subsumed into step 1:

1’. Make it happen. Select the objects that will be included in the study and make the phenomenon appear that you intend to study.

The separation step is crucial in a scientific experiment. There are two major ways to do this, elimination and effect separation (Hansson, 2007). R.A. Fisher stressed that in controlled experiments the purpose of introducing randomness into the design is to separate systematic variation from purely random error (Fienberg and Tanur, 1987, p. 77), that is, to conduct the separation step of the experiments. By ‘control’ Hansson refers to the control of external factors that may make the experiment inefficient. A standard procedure to control those is blocking.

3 What does an experimental statistician work with?

It is useful in research to distinguish between extraneous sources and what Kish (1987, pp. 3-5) calls explanatory variables, which comprise predictors
(independent variables) and predictands (dependent variables). The extraneous sources are largely undesirable in research. Kish classifies variables that come from extraneous sources into three groups, the two first being:

1. Controlled variables. “Control may be exercised either by design of the selection procedures or by estimation techniques in the statistical analysis, or perhaps by both” (Kish, 1987, p. 4).

2. Disturbing variables. These are confounding variables.

Kish (1965, p. 594) writes that the aim of efficient design, both in experiments and in surveys, is to move as many extraneous variables as is feasible into the class of controlled variables, and to control for them successfully. A common way to exercise control (i.e. move variables from being disturbing to the class controlled) through estimation is regression analysis. For example, if a researcher discovers that the incidence of cancer is higher in Bournemouth than in England in general, he can include age among the predictors and the otherwise disturbing variable age is ‘controlled for’ (and the higher incidence in Bournemouth will disappear). Various kinds of regression analyses are common also in survey sampling (Särndal et al., 1992).

Consider a field in an agricultural experiment. The soil will vary in all directions and hence potentially disturb the inference. How can we define blocks and find explanatory factors that accounts for all disturbing factors? Salsburg (2001, p. 47) paints the picture of how a discussion in Rothamsted may have gone:

R.A. Fisher sits back and smiles, letting the other researchers get more and more involved in complicated constructs. He has already considered these questions and has a simple answer. Those who knew him describe Fisher as sitting, quietly puffing on his pipe, while arguments raged about him, waiting for the moment when he could insert his answer. He removed the pipe from his mouth.

Randomise, he says.

Randomisation is an ingenious trick that helps to move variables from the class ‘disturbing’ to the class ‘randomised’. So added to controlled variables and disturbing variables above, Kish has a third group:
3. Randomised variables. They are uncontrolled but treated as random errors.

Kish (1987) notes in his Figure 1.2.1 that in a successful and well-designed experiment all disturbing variables have been transformed to controlled or randomised variables.

Unlike the (successful) experimental statistician, the survey sampler will have to address remaining disturbing variables that for example nonresponse, measurement error and undercoverage have created.

4 What does a survey sampler work with?

The steps 1’, 2, 3, 4 in the section ‘What is an experiment?’ are actually the same in a survey, with one exception. What does a survey practitioner do when conducting these steps?

1. Select the objects that will be included in the study and observe/measure the phenomenon/characteristic that you intend to study.
2. Separate. Disturbing factors needs to be removed.
3. Control. Factors that may influence the phenomenon must be controlled.
4. Observe. Obtain as exact observations as possible.

It is not usual in survey sampling to instigate a phenomenon; rather the survey practitioner strives to observe or measure without exerting influence on the phenomenon that is the object of the survey. Apart from this, steps 1 to 4 are at an abstract level similar in surveys and experiments. However, at a more practical level they are quite different:

1. The selection of units is a step that the survey practitioner devotes ample time to. The thinking about the best way to select units starts with a definition of the target population, which is usually a finite set. Routes to find units in the target population and observe their study variables are identified or devised.
2. Disturbing factors in surveys are removed or ameliorated. To give an example, interviewer effects are undesirable but unavoidable in telephone or face-to-face interviews (Biemer and Lyberg, 2003, Ch. 5). They are mitigated through training and clear procedures.

3. Control is exercised through sampling design (e.g. stratification) or estimation (e.g. generalised regression estimation, Särndal et al. (1992)).

4. The observations are made through a mode (any medium that is used to measure or collect information from observed units). Biemer and Lyberg (2003, Ch. 6) give an excellent account on modes and their (disturbing) errors.

As noted above, disturbing factors in surveys abound. As early as in the beginning of the 20th century, Bowley addressed the measurement problem in his survey on daily household income among working-class families in Reading (Bowley, 1913). He used a model to compute weekly household income.

5 How is randomness created?

The statistician has to pinpoint the source of randomness she or he is going to take into account. This is actually an intersubjective procedure. In some areas of statistics randomness is trust upon the objects of study. Design-based survey sampling and randomised experiments are two of those areas. Consider a target population of \( N \) objects, \( U = \{u_1, u_2, \ldots, u_N\} \), and an indicator \( I = \{I_1, I_2, \ldots, I_N\} \). In a survey,

\[
I_k = \begin{cases} 
0 & \text{if unit } k \text{ is not included in the sample} \\
1 & \text{if unit } k \text{ is included in the sample}
\end{cases}
\]

In an experiment with two treatments,

\[
I_k = \begin{cases} 
0 & \text{if unit } k \text{ is given treatment A} \\
1 & \text{treatment B}
\end{cases}
\]

The purpose is to observe a variable \( y = \{y_1, y_2, \ldots, y_n\} \) in the sample of size \( n \). The variable \( y \) is referred to as study variable in survey sampling and response variable in experiments. It can be multivariate.
In both surveys and experiments there is a set $S$ of all vectors $I$. In both cases, $S$ is the set that defines the randomness which helps the statistician to place the actual result of the survey or experiment in a certain position of the stochastic distribution and make inference.

In surveys, a sample is one realisation of the stochastic vector $I$. In design-based survey sampling, a sampling design is a function $p(\cdot)$ such that $p(I_s)$ is the probability of selecting the sample $I_s$. For example, simple random sampling is defined as the function $p(\cdot)$ where all $p(s)$ are the same and all samples have the same predetermined size, $n$. The probability $p(s)$ in simple random sampling is the inverse of $\binom{N}{n}$. The probability of selecting one particular unit is referred to as inclusion probability. Unfortunately, simple random sampling is often referred to as ‘random sampling’. The term ‘random sampling’ for simple random sampling is cursory, or even sloppy, as there are many other random sampling designs. In fact, research into different random sampling designs was intense in the middle of the 20th century. Around 1980 the focus changed to estimation.

The design-based framework was founded by Bowley (1913) and Neyman (1934).

Random sampling plays a less critical role in model-based survey sampling. To the contrary, it has been argued that random sampling is not only unneeded, it may even be counter-productive (Valliant et al., 2000, p. 19). In model-based survey sampling a large number of tentative samples are drawn and the sample that is most ‘balanced’ is selected for data collection. The most basic form of balance is $\bar{x}_s = \bar{X}_U$, where $x = \{x_1, x_2, \ldots, x_N\}$ is a variable known for all units in the target population and $\bar{x}_s$ and $\bar{X}_U$ are the sample and population averages. Randomness is not devised by the statistician, it is rather assumed to be inherent in the study variable that is assumed to have some stochastic distribution. Model-based survey sampling is of a more recent date than the design-based framework, despite the fact that it is in some respects closer to mainstream statistics than is design-based survey sampling. The two articles introducing model-based survey sampling are Brewer (1963) and Royall (1970).

The main difference between experiments and model-based survey sampling in terms of the role of randomness is that in experiments treatment is randomised to separate factors (see the section ‘What is an experiment?’) whereas in model-based survey sampling there is no active randomisation. In
both cases, however, randomness constitutes ‘left-overs’ when everything that is feasible to control has been addressed. This randomness stems from the perceived or real stochastic nature of $y = \{y_1, y_2, \ldots, y_N\}$. In design-based survey sampling the study variable is believed to be, or taken as, non-stochastic.

6 What drove the separation?

One clue to the root of the separation of surveys and experiments can be sensed in the very first sentence in Stephan’s paper on the history of survey sampling (1948, p. 12): ‘Modern sampling practice is the result of combining statistical theory with scientific knowledge about the material that is being sampled’ (my emphasis).

In conversation (Fienberg and Tanur, 1987, p. 86) William Madow noted three factors driving the separation of experiments and surveys:

1. the complexity of the populations sampled in large scale surveys (heterogeneity, skewness, and mixture properties)

2. the large sizes of samples selected in large scale surveys made it possible to draw inferences that did depend on a probability structure imposed by the survey designer and did not depend on assumed probability densities (see design-based and model-based survey sampling above)

3. from the early work of Fisher the simplest analysis of variance model did not permit a negative intraclass correlation coefficient, while cluster sampling as defined for finite sampling would yield a negative intraclass correlation

The survey samplers around the 1940s, William Madow included, addressed the issue of population complexity by introducing sampling designs with unequal inclusion probabilities. Knowledge about the population was essential to devise a cost-efficient sampling design. There was an enormous growth of complex sampling designs. For example, Brewer and Hanif (1983) list more than 40 sampling designs of the $\pi ps$ type (probability proportional to size without replacement) that they had found in the literature. Some modern textbooks, for example Chambers and Clark (2012), are organised around
types of population, with chapter headings such as ‘homogenous population’, ‘clustered population’, etc.

The textbook by Hansen et al. (1953) played an instrumental role in shaping the survey sampling practices. They include fairly little material on randomised experiments. That is in fact true for the vast majority of textbooks in surveys sampling with one notable exception: Yates (1981). Frank Yates was early in his career employed by R.A. Fisher in Rothamsted.

Another driver of the separation was the growth of the literature in both fields. It just became very hard for newcomers to statistics to master both survey sampling and randomised experiments (Fienberg and Tanur, 1987).

7 Different desiderata

You cannot have it all in research. It is sometimes useful to map the aim of research onto three dimensions: representation, control and realism (Kish, 1987; O’Muircheartaigh, 1999). These can be called desiderata (Weisberg, 2006). By representation we mean whether it is possible to generalise to a wider population, by control we mean whether confounders are in control and finally by realism we have the validity of the research in mind.

Kish (1987, p. 9) notes that only rarely all three desiderata can be satisfied adequately in any research, and often you can have only one of the three. Kish goes on to say that experiments are strong on control, weak on representation and often also on realism. Perhaps Kish had in mind the type of experiments where subjects are chosen out of convenience. For example, Godden and Baddeley (1975, 1980) test the effect of changing the physical context on how well subjects recall a list of words. The subjects were divers from the university diving club and the physical contexts were under-water and above water. In Phase III clinical trials the sampling design is certainly not a convenience sample but still it may fall considerably short of what in official statistics would be considered as a good, random sample that provides reliable grounds for representation. So there is more than a grain of truth in the claim that experiments are weak on representation.

Surveys, Kish (1987) claims, are strong on representation and realism. They are strong on realism because “measurements can often be made in the
‘natural settings’ of actual populations” (Kish, 1987, p. 9). I disagree with
the last point. I believe that collecting data through a survey is very different
from making real-life observations. An exception may be surveys on crop
yields and similar topics.

A fourth desideratum may be precision. Surveys often sacrifice precision
in study variables to be measured in the survey. One common sacrifice is the
use of Lickert scales or similar where something that is more or less continuous
is mapped onto a, say, five-point scale. Another sacrifice is when the survey
designer deliberately simplifies questions (for example, by asking how often
respondents eat vegetables instead of asking about the amount of vegetables).
In sum, I believe surveys are strong on representation only, although in terms
of control there have been substantial recent advances in estimation where
‘auxiliary’ variables (controlled variables) remove disturbing variables, not
only to reduce bias but also to improve precision in estimates.

8 Exploring the rationale behind the separa-
tion

Different origins and lack of communication aside, are there rational reasons
for the fact that design in survey sampling and design of randomised experi-
ments are separate areas of statistical science?

Our notes above suggest four possible reasons:

1. Experimental research and survey sampling may have different aims.
2. Survey sampling faces substantial challenges with disturbing variables.
3. Knowledge about the population structure is paramount in surveys,
   which has led to an interest in complex sampling designs.
4. In the survey sampling area there is a choice between design-based
   inference, where moments are taken over \( p(\cdot) \), and model-based in-
   ference where moments are taken over the non-sampled part of \( \mathbf{y} = \{y_1, y_2, \ldots, y_N\} \) conditional on \( I_s \).
To the discussion above on different aims (realism, precision, control and representation) the dimension analytical and descriptive aim should be added (Thompson, 1997). You have a descriptive aim when you strive at estimating a function of $y = \{y_1, y_2, \ldots, y_N\}$ (uni- or multivariate), and an analytical aim when the goal is to estimate parameters in a model that captures the relationship between predictors and predictands. Surveys have usually descriptive aims and experiments analytical aims.

A further intriguing issue, which we have not touched upon so far, is whether survey samplers and experimental statisticians have different ways of thinking about optimality in terms of inferential precision. This issue can be organised at least under the three headings: ‘interest’, ‘what kind of critical points’ and ‘what objective function (loss function)’.

To start with ‘interest’, the mere existence of a book like Atkinson et al. (2007) suggests a stronger emphasis on optimality in experiments. However, in practice, experimental designs are often chosen in a traditional fashion, even in expensive clinical trials (Fackle-Fornius et al., 2015). Optimality plays a backstage role in surveys. Results on optimal choices like Neyman allocation (Neyman, 1934) appeared rather early in the development of survey theory, but now cost-efficiency, rather than optimality, attracts greater interest. The main reason may be the fact that most surveys have several purposes: there are many study variables and subpopulations. In model-based sampling the focus shifted early from optimality to robustness, the reason being fear of bias (Valliant et al., 2000, Ch. 3).

Turning to ‘what loss function’, mean squared error is the by far most often used loss function in surveys. I would argue that at least in official statistics the most relevant function is the ‘survey error’, that is, the difference between the true finite population parameter and the estimate. The central position of the MSE appears to be due to its mathematical tractability.

Finally, as for the issue ‘what kind of critical points’ we note that experimental design with nonlinear models and surveys share one obstacle: the optimum depends on unknown parameters and hence the optimum is strictly speaking infeasible to identify ahead of the data collection. Practitioners use estimates (‘guestimates’) of those parameters to compute an optimal design in terms of minimum MSE. Instead of trying to locate a minimum of the objective function, a minimax type of criterion to find a good (minimum of the maximum variance over a reasonable range of values of the unknown pa-
rameters) optimum seems preferable (Scott and Smith, 1975; Cheng and Li, 1987; Fackle-Fornius et al., 2015). Hedlin (2002) suggests minimum risk of obtaining estimates with large error as a criterion. The downside is loss of mathematical tractability.

The interesting issue whether optimality is an area where survey sampling and randomised experiments can learn from each other is left for future research.

This article started with a quotation from Kish (1965): “The separation of the sample surveys from experimental designs is an accident in the recent history of science... ”. It is intriguing to see what follows that quotation:

“... and it will diminish. With complete command over the research situation, we could introduce, with firm experimental controls, the desired effects and measurements into controlled and randomized portions of the entire target population”.

References


