Comments on the development of Survey Statistics theory and practice in the last fifty years; some personal reflections.

Carl-Erik Särndal

Statistics Sweden

Abstract

The continued development of the field of Survey Statistics will be interesting. One reason is that the data collection phase for producing official statistics is likely to change, possibly to alternatives other than the probability sampling data collection that has been a standard, or seen as an ideal. This presentation cannot predict the development; it looks instead at some important ideas in the progression of Survey Statistics over the last five decades.

In the more than one hundred years history of survey sampling, a more than sixty year old result has had a particular significance, namely, that unbiased estimation of a finite population total is obtained by weighting the observed survey variable values by the inverse of the inclusion probabilities. This works, because in probability sampling, these probabilities are known for all population units. The unbiased estimator that expresses this fundamental and mathematically simple result bears the name of the two auteurs, Horvitz and Thompson, of a classical 1952 JASA article. But behind the result lay a long development, from the early attempts of statisticians to convince users of statistics that "observing just a small sample from the large finite population can be enough". Although interesting in a historical perspective, this long period is not considered in this presentation.

Inverse inclusion probability weighting, and its modifications and extensions, have had a strong impact on survey statistics over the last fifty years, which is the period examined here. Such weighting is the basis for what we now call *design-based inference*.

By contrast, an alternative approach known as *model-based inference* will, at least in its most pure forms, deny that any important role be given to probability sampling and to inverse inclusion probability weighting. Modeling, and trust in the assumed models, is the justification for the inference. Although not design unbiased, the resulting estimates may be advantageous in other ways.

A feature of the last fifty years of development is the importance of auxiliary variables in

the estimation process for official statistics. This has been particularly evident in northern European countries, with their access to a vast supply of auxiliary variables, from administrative registers, or in the form of paradata.

Several areas of research and practice have extended the design-based inference paradigm. The two areas mentioned below were, in their original form, presented for a survey background that national statistical institutes cannot count on now, several decades later: a full, or almost full, 100% response from the selected probability sample.

1) The generalized regression (GREG) approach originated in the realization that while inverse inclusion probability weighting is needed for design unbiased estimation, such unbiasedness is not the only important factor. The estimation also needs to be variance efficient. The GREG estimation approach realizes a low variance from a strong regression existing between survey variable y and auxiliary vector $\mathbf{x}$. One can explain it by saying that accurate prediction of the unobserved y-values is derived from the information on $\mathbf{x}$ known for the population.

2) The calibration (CAL) approach had its origin in a search for a weighting of the observed sample y-values that is not far from the basic inverse probability weighting, but better than it, because required to respect a condition called a calibration equation, where the known population total of the x-vector, or a design unbiased estimate of it, appears on one of the two sides of the equation. But a secondary purpose is to explain the survey variable y through the auxiliary vector x. The calibration approach is thus double-natured: The weighting aspect is combined with an implicit relationship between y and $\mathbf{x}$. Although the outlook is different, the CAL approach is in special cases identical to the GREG approach.

Both 1) and 2) can be called *design-based model assisted inference*. However, the last few decades have witnessed a strong adverse trend for the conditions for probability sampling surveys: high rates of nonresponse in the drawn probability sample. It has become necessary to adapt the inference – which can perhaps no longer be called design-based - to these new conditions.

High nonresponse causes a more or less pronounced bias in the survey estimates. This can happen even under conditions of quite strong relationship between y and $\mathbf{x}$. The objective is then to hold this bias as low as possible. The CAL approach that has been particularly important and useful for nonresponse weighting adjustment. Auxiliary variables are also important for managing the data collection so as to get a well-balanced set of respondents from the drawn probability sample.

The presentation reviews briefly the approaches 1) and 2), then focuses on approaches to inference under (high) survey nonresponse. A question arising is: How important will the probability sampling paradigm and the inverse inclusion probability weighting be in the future?