Population Review

Volume 46, Number 2, 2007

Type: Book Review pp. 63-65

Indirect Sampling

Author: Pierre Lavallée
Publisher: Springer, New York (2007)
Pages: 245
ISBN: 978-0-387-707785

Reviewer: Gebrenegus Ghilagaber
Affiliation: Department of Statistics, Stockholm University
Corresponding author/address: Gebrenegus Ghilagaber, Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: (Gebre@stat.su.se)

The book has its roots in a doctoral dissertation that was completed in June 2001 at the *Université Libre de Bruxelles*, Belgium. The dissertation, perhaps a revision of it, appeared as a book in French, *Le sondage indirect, ou la méthode généralisée du portage des poids*, which was published simultaneously by *Éditions de l'Université de Bruxelles* (Belgium), and *Éditions Ellipse* (France) in 2002. The current book is an English-translation of the French version with 'some sections added to reflect the new developments that have occurred since 2002'.

In the more common (direct) sampling, a random sample of individuals is drawn from lists known as sampling frames representing the target population – the set of individuals under investigation. In such a case, individuals are selected with known probabilities of selection and these probabilities can be used to obtain the reliability of the information produced by the survey.

The need for indirect sampling arises when there is no sampling frame for the target population – a situation that, according to the author, is very common even in National Statistical Offices. Thus, the main motive of the book is to propose and develop techniques of getting valuable information and reliable estimates on the target population with no access to it but using a sampling frame of a population that is related to it.

A typical example is when one is interested to conduct a survey on children. Since a sampling frame of children may not be available one has to resort to adults - a population related to the target population (children) and for which sampling frame is available. Thus, a sample of adults is drawn from the available sampling frame, and all the children of the selected adults are surveyed. Here, a tacit assumption is that all children have at least one selectable parent. The book also provides a second example where indirect sampling may be more problematic: when only a list of businessmen and businesswomen is available a survey of businesses becomes complicated because a business person may own more than one business or that a business may be owned by more than one person.

The book is divided into ten chapters presenting theory, algorithms, and applications; as well treating difficulties brought by auxiliary information, non-response, or by the combination of different sources using record linkage.

The introductory chapter begins with a brief description of sampling theory and cluster sampling, and introduces indirect sampling, together with the associated weighting method – the Generalised Weight Share Method (GWSM). The GWSM is described at greater length in Chapter 2 and its merits are

highlighted in connection with indirect sampling for rare populations, weighting using only selection probabilities of the selected units, weighting of populations related by complex links, and weighting of unlinked units. Chapter 3 is devoted to a brief review of the literature related to the foundations of GWSM – including Fair Share Method and its generalisation, Network Sampling, Adaptive Cluster Sampling, and Snowball Sampling.

Properties of the estimator based on GWSM are outlined in Chapter 4. It is shown that the estimator is unbiased and a formula for its variance is provided together with a method for its estimation. Further, it is shown that under conventional cluster sampling the GWSM estimator yields the same results as those obtained in classical theory. Since the GWSM estimator of the total is not guaranteed to have the smallest variance, the chapter ends with a description of a procedure for variance reduction. The procedure is based on obtaining optimal weighted links and then applying the well familiar Rao-Blackwell theorem – conditioning an unbiased estimator on a sufficient statistic in order to obtain a new estimator with at most equal variance as the unbiased estimator.

The GWSM estimator may be viewed as a generalisation of weight share method, network sampling, or snowball sampling. It is further generalised to two-stage sampling in Chapter 5, which also includes a discussion on the arbitrary aspect of cluster-formation and possibility of eliminating the notion of clusters. An application to longitudinal surveys is outlined in Chapter 6. In Chapter 7, the GWSM is associated with Calibration in order to incorporate auxiliary variables with a view to improving the precision of GWSM estimates.

Non-response in indirect sampling is treated in Chapter 8 where different techniques to take due account of non-response are presented. The techniques are based on adjusting the estimation-weights obtained by the GWSM and, thus, exclude techniques that are centred on the imputation of missing values. In Chapter 9, it is shown that GWSM can handle estimation problems related to data from different sources combined using record linkage. This is achieved by adapting the GWSM to account for linkage weights obtained from record linkage.

The final and concluding chapter spells out the main strengths of the GWSM, lists some other works that made use of GWSM in diverse applications since its introduction by the same author in 1995, and highlights potential future works in indirect sampling in general, and GWSM in particular. The applications that drew on GWSM to solve concrete problems associated with indirect sampling include (i) selecting establishments in order to survey enterprises having these establishments in Canada, (ii) selecting a set of services provided to homeless at service centres in order to survey homeless persons in France, (iii) a survey of tourists in France that is built from three different frames: a subset of the most visited attractions, highway payment polls, and a sample of bakeries, (iv) selecting postmen in order to estimate the flow of mail (envelopes, packages, etc.) at the French National Mail Agency, and (v) using GWSM to weight a sample of towns for the estimation of social security beneficiaries in Switzerland.

In reading the book one learns that indirect sampling is a form of probability sampling and that one can get precision of the resulting estimates. However, computation of these selection probabilities becomes complicated since, in indirect sampling, sampling frame and target population are distinct. The GWSM exploits the relationship between target population and sampling frame in estimating the selection probabilities.

For the benefit of this journal's readership this reviewer would like to comment on how data collected through indirect sampling would be analyzed subsequently. If, for instance, to conduct a survey of children, a sample of adults is drawn from available sampling frames, and all the children of these adults are surveyed, then it is only natural that the children are nested within their corresponding parent. This is the case in Demographic and Health Surveys or Fertility Surveys to which readers of this journal may be familiar. Since children of the same parent are more alike than children selected at random from the population, the basic assumption of a random (independent) sample of children can't be ascertained. Thus, in any subsequent analysis of children-data that is collected through indirect

sampling, analytical methods must pay due attention to the hierarchical nature of the data. This may be achieved by treating children from the same parents as correlated cases (multi-levels) within the same observation (parent) and use some form of multilevel modelling that are now abundant in the literature. Such a procedure also has the additional advantage of accounting for any parent-specific unobserved heterogeneity that may affect outcome at a child level. While data collected through cluster sampling are hierarchical (multilevel) it is not clear, at least not explicitly to the present reviewer, whether the weighting system in GWSM addresses the issues related to subsequent analysis of data collected through indirect cluster sampling.

So, how would this book fare in the academic and scientific community and who can benefit most out of it? The book does not clearly define the readership it is intended for or the level of mathematics required to follow it without much strain. Since it originates from a doctoral dissertation, it is understandable that the focus is on one particular method – Indirect Sampling and the associated weighting system, GWSM – though attempt is made to link it to most of the classical methods. This, together with the lack of illustrative examples and exercises, makes it less ideal as a basic text for a course in the applied sciences that is based on classroom teaching.

On the other hand, it is a welcome injection to the literature on survey methodology and sampling theory. It is certainly a valuable reference to researchers and professionals in survey methodology. Those familiar with the formula and well versed in the mathematics the book demands can draw maximal benefit out of it.

As his dissertation-supervisor writes in the foreword, Pierre Lavallée is among the few colleagues who can boast to posses the recognized competence in the domains of both Sampling Theory and Survey Practice and this reviewer commends the author for the work well done.