Recent Developments in Sample Survey Theory and their impact on Official Statistics

T.M..Fred Smith
University of Southampton, Department of Mathematics,
Southampton, SO17 1BJ, U.K.

1.Introduction

It is a great honour to be invited to give this lecture to commemorate the 25th anniversary of the IASS. In 1971 the ISI Council voted to form a new section, the IASS, and the first sessions organised by the IASS were at the Vienna meeting in 1973The proceedings of the IASS have been published since 1975, and we are now into edition 39 of the Survey Statistician. 1975 also saw the first publication of Survey Methodology, while 1985 saw the launch of the Journal of Official Statistics. Survey sampling now has several major outlets for the dissemination of theoretical results and discussions of practical issues. In reviewing recent research as reflected at the IASS several themes recur throughout the entire period; censuses, sampling errors and their computation, consumer price indices, establishment and household surveys, market research, the presentation of official statistics, and non-sampling errors of all forms. There are also themes that develop in importance; edit and imputation, co-ordination of surveys, small area estimation, repeated surveys, and the use of computers. Themes that emerge include; confidentiality, computer assisted interviewing, multiple imputation, statistical indicators, time series methods, quality assurance and total survey error. A theme that declines is that of the foundations of finite population inference.. It is frequently asserted that there is a divide between sample survey theory and sample survey practice. The fact that the assertion is made means that there is a problem, but I am going to argue that most good survey practice is grounded in theory. Another divide that existed in 1973 was between sample survey theory and the rest of statistical theory. My thesis is that both divisions have been largely closed.

When in 1925 the ISI finally accepted the case for the use of representative samples in official statistics it was because the report proposed both a method for selecting representative samples, simple random sampling, and also a theory for measuring the uncertainty due to sampling. Without this framework representative sampling lacked credibility. It was difficult to extend this theory to more complex sampling schemes and the next major advance was by Neyman (1934) who changed the theoretical basis of survey inference from the hypergeometric likelihood function to sampling errors based on the randomisation distribution. From then theory and practice developed rapidly together under the leadership of people like Morris Hansen, the first president of the IASS. By the 1950s randomisation theory was almost complete and attention switched from sampling errors to non-sampling errors, again under Hansen's leadership. The contributions during this period were based on a coherent theory developing from practical problems.

In the 1950s and 60s theoreticians addressed the foundations of randomisation inference. Attempts to integrate randomisation inference into mainstream inference, for example, Godambe (1966), Basu (1971), were largely negative and this led to a search for an alternative model-based framework for survey inference. This work seemed abstract to practitioners and may have resulted in the perceived divide between theory and practice. Theoreticians, such as Ericson (1969), Scott and Smith (1969) and Royall (1970), showed that a modelling approach could be adapted to complex finite population structures and sampling schemes. For a review see, for example, Smith (1976). When we pick up the story in 1973 one of the big issues is randomisation versus model-based inference.

2. Model-based inference and randomisation inference

Given the difficulty of inductive inference it is not surprising that there should be alternative approaches. At the Vienna ISI meeting Fuller (1975) proposed a set-up for the regression analysis of survey data, while Brewer and Mellor (1973) helped to clarify the issues through a dialogue between Harry, a practical survey statistician, and Fred, who is younger and more theoretically inclined. The first clarification is the recognition of the importance of defining the target population for inference. If the target is the fixed finite population from which the sample was drawn then the inferences are descriptive. If the target is some other population, so that even a perfect census leaves inferential uncertainty, then the inference is analytic, a term introduced by Deming (1950). For analytic inference models are a necessity, but for descriptive inference there is a choice between model-based predictive inference and randomisation inference.

An argument for randomisation inference is that since it does not depend on models it is robust. Smith (1994) shows that the argument lacks substance and can be explained in terms of a transfer between bias and variance. Also randomisation inference is not assumption free, it depends strongly on the normal approximation. It is easy to construct examples where the normal approximation fails, especially when the population has outliers. Cochran (1977) gives a rule for the minimum sample size necessary for the normal approximation to hold for a standardised statistic under SRS. Sugden et al (1999) extend this to Studentised statistics. If G_1 is Fisher's measure of skewness then they propose that $n>28+25G_1^2$. Outliers lead to large values of G_1 and even in large surveys the sample size in domains of study may fail this condition.

Royall (1976) argued that if you know the values of certain covariates in both the sample and the population then inferences should be based on samples like the one actually selected and not on the distribution of all possible samples. In Fisherian terms the sample belongs to a relevant subset. Model-based inferences condition on the covariates and the sample units selected, and Royall demonstrated empirically that model-based interval estimates could have far better conditional coverage properties than the corresponding randomisation intervals. This presented a serious challenge to advocates of randomisation inference, see Holt and Smith (1979), Rao (1985). A new approach was needed and Robinson (1987) provided it for the ratio estimator by employing the asymptotic normality of the joint distribution of the survey variable, Y, and the covariate, X, to derive the conditional distribution of Y given X, and hence a form of conditional randomisation inference. Casady and Valliant (1993) use this approach to construct conditional inferences for post-stratification. This is an area where new theory, motivated by a seemingly esoteric debate, may lead to a change in practice with regard to the way sampling errors are presented and interpreted.

Royall's model-based analysis also showed that some samples are more representative than others. In this case why rely upon the "on the average" properties of random samples when it is possible to select purposive samples balanced for the known covariates? When one considers the multiple aims of most surveys, the many different target populations and the myriad variables studied, then the case for the "on the average" protection of random sampling, rather than dependence on balance within a class of models, becomes compelling to most statisticians, see Hansen, Madow and Tepping (1983). Randomisation in design is a defensible strategy and approximate balance can be achieved by stratification. But exact balance over covariates is also appealing and this is captured in recent work on calibration, Deville and Sarndal (1992), which is balance in estimation as opposed to balance by design. I conclude that some of the ideas in the theoretical debates about foundations are now being absorbed into practice, albeit in ways not foreseen at the time.

3. Combining models and randomisation

Clogg and Dajani (1991) identify six processes to consider in finite population sampling; the generation of the finite population values (the super-population model), the generation of the sample, the contact and measurement processes, summarising complex data and the process of inference. Statisticians should report on all of these processes. This requires an integration of model-based methods into sample survey inference since only the sampling process has a strict randomisation justification. In practice total survey error is systematically under-reported by concentration on sampling errors. It has long been recognised that non-sampling errors are at least as important as sampling errors, and that their relative importance increases with sample size. Despite the model-based work of Hartley and Rao (1978) who showed how to design surveys to measure several components of survey variance simultaneously, and hence to estimate total survey variance, there have been no published studies where this has been attempted. Variances are only part of the story and non-sampling biases can invalidate inferences and dominate the overall error when the sample size is large. Linacre and Trewin (1993) is a rare example of a case study that considers several components of total error, including both biases and variances, in order efficiently to redesign an establishment survey. Practitioners should stop playing lip service to total survey error and use existing theory to design studies where all the components are measured so that the ideas of total quality assurance, Deming (1986), can be implemented.

A major advance in controlling non-sampling errors has been the use of computer assisted interviewing which was developed by methodologists, such as Bethlehem and Keller (1991), from a desire to control the overall survey process.. This has dramatically reduced the need for, and cost of, editing. CAI, and in particular CATI, presents opportunities for on-line experimentation which have yet to be fully exploited. CAI helps to reduce missing items, but not missing units. Little and Rubin (1987) clarify the assumptions underlying the missing value process and propose model-based methods for handling them, including multiple imputation, Rubin (1987), which captures the additional uncertainty due to imputation. CAI controls some elements, but the questionnaire remains the weakest link in the measurement process. Is a standardised questionnaire the best way to elicit accurate responses from widely differing units? Are there any theories that will help us?

Official statisticians are frequently asked for information on small areas for the purposes of resource allocation. Effective sample sizes are usually too small for accurate direct estimation and so methods have been devised for borrowing data from similar areas. Mixed linear models capture the implicit assumptions and enable variances to be estimated, see Ghosh and Rao (1994). However, statisticians should acknowledge the limitations of these methods which depend strongly on the model assumptions and the availability of covariates. One alternative is to use administrative data collected specifically for this purpose, another is to increase sample size. Both are costly, but the clients should be forced to face up to the consequences of their requests for information.

Most official surveys are longitudinal and the output is a time series of cross-sectional totals or means, Kalton and Citro (1993). Randomisation inference treats the population totals as unknown constants, not as a time series. Scott et al (1977) suggested that they should be modelled as the realisation of a time series with superimposed sampling error. There are many possible approaches and the Kalman filter has proved to be a fruitful method for analysing time series of survey estimates, Pfeffermann and Bleuer (1993). There have also been major theoretical advances in the analysis of longitudinal micro-data which have been incorporated into sample survey analysis, see Skinner (1998). Another major advance has been in the use of hierarchical models to capture the context of social processes, Goldstein (1995). Together with other advances in analytic surveys, Rao and Scott (1984), Skinner et al (1989), the adoption of these new methods has helped to bring surveys into mainstream statistics.

When the finite population is modelled as a sample from an infinite super-population the target parameters can be defined using population level estimating equations, Godambe and Thompson (1986), Binder and Patak (1994), which can themselves be estimated using the sample

inclusion probabilities. This use of estimating equations is an example of a model-assisted approach to finite population inference as exemplified in the 'yellow book', Särndal et al (1992). Unbiased estimating equations lead to the Hajek version of the Horvitz-Thompson estimator, which has many desirable properties. Practice leads theory here, since the standard HT estimator is rarely used in practice.

Traditional randomisation theory has continued to develop. Chao (1982), Tillé (1996) have devised πps sampling schemes that enable joint inclusion probabilities to be easily calculated. Exact sampling variance estimation is still the exception and most work on variances has been based on re-sampling methods such as the jackknife, BRR and the bootstrap, see Rao (1998). These methods capture some elements of non-sampling variance but we still await implementation of random group designs that will enable total variance to be estimated, Mahalanobis (1946), Deming (1950).

The above developments have been mainly by theorists tackling practical problems in the best statistical tradition. There has also been a reconsideration of the nature of statistical models, on which so much theory depends, reflecting a healthy interaction between theory and practice. Best practice now combines traditional sampling theory with model-based ideas in a fruitful manner.

Emerging trends are driven by the increase in computing power and the insatiable demand from users for more information. Felligi and Wolfson (1997) highlight the need for a coherent system for social on a par with national accounts. The same is true for environmental statistics. We need criteria for the construction of summary measures for monitoring policy outcomes based on complex multivariate data. The merging of data from different sources and the construction of data bases which combine administrative, survey and qualitative data while maintaining confidentiality is already challenging theorists and practitioners. We also need to develop a general theory for statistical description that is not based on models. Above all we must recognise that surveys have limitations and should not be used to answer questions for which they are unsuited..

REFERENCES

The following list is of books and review articles that contain references to most of the cited papers. Others may be found in the proceedings of ISI meetings, JOS and Survey Methodology.

Cochran,W.G. (1977). Sampling Techniques, 3rd ed. New York: Wiley.

Deming,W.E. (1950). Some Theory of Sampling. New York: Wiley.

Ghosh,M.and Rao,J.N.K. (1994). Small area estimation: an appraisal. Statist. Sci., 9, 55-93.

Kalton,G. and Citro,C.F. (1993). Panel surveys: adding the fourth dimension. Survey Methodology, 19, 205-215.

Little,R.J.A. and Rubin,D.B. (1987). Statistical Analysis with Missing Data. New York: Wiley.

Rao,J.N.K. (1998). Some current trends in sample survey theory and methods. Manuscript.

Rubin,D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

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

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

Smith,T.M.F. (1976). The foundations of survey sampling: a review. J. Roy. Statist. Soc. A, 139, 183-204.

Smith,T.M.F. (1994). Sample surveys 1975-1990; an age of reconciliation? Int. Stat. Rev., 62, 3-34.

RÉSUMÉ

Cet article revue les développements des techniques d'enquête depuis 1973.