THE PAST, PRESENT AND FUTURE OF SURVEY SAMPLING

by

Professor Romanus Odhiambo Otieno

July 23, 2012

Outline of Presentation

- Background of Survey Sampling.
- Approaches to Survey Sampling.
- Various broad areas of Survey Sampling.
- Survey Sampling to date.
- Prospects of Survey Sampling

Definition of Survey Sampling.

This area of Statistics is concerned with selecting subsets of the units (samples), observing features of the sample units, then using those observations to make inferences about entire populations.

Sampling Theory is distinguished from the rest of Statistics by its focus on the actual population of which the sample is a part.

In other areas of Statistics, observations are typically represented as realizations of random variables, and the inferences refer not to any actual population of units, but to the probability law that governs the random variables.

Problem of Survey Sampling

Suppose that the number of units N in the finite population is known and that with each unit, is associated a number y_i.

The general problem is to choose some of the units as a sample, observe the $y's$ for the sample units and then use those observations to estimate the value of some function, say,

$$h(y_1, y_2, y_3, \ldots, y_N)$$

of all the $y's$ in the population. The function

$$h(y_1, y_2, y_3, \ldots, y_N)$$

can be a simple combination of the $y's$ like their total or their mean or may even be something more complex like a quantile.

To perform a prediction, we use the popularly known Prediction Approach to Survey Sampling by treating the numbers

$$y_1, y_2, y_3, \cdots, y_N$$

as realizations of random variables

$$Y_1, Y_2, Y_3, \ldots, Y_N$$

After the sample has been observed, estimating $h(y_1, y_2, y_3, \ldots, y_N)$ is about predicting a function of the unobserved $Y's$.

Usually, the relationships among the random variables are expressed in a model for their joint probability distribution, and predictions are made with reference to this model.

For example, a sample of parts coming off an assembly line could be tested to determine how many meet engineering specifications.

A Statistician might then represent the results probabilistically, with each part having the same unknown probability θ of being defective, and seek to estimate θ or test whether it is less than some acceptable level.

The problem becomes one of sampling if attention shifts from the probability θ to the actual proportion of detectives in a day's production. An interested reader may see Anthony [1], Bashtannyk and Hyndman [3], Beaumont and Bocci [4] amongst others.

Background of Sampling Theory

Survey Sampling as an area of Statistics has been mainly theoretical and only started to appear in the 1940s.

Detailed research started with the work of Armitage [2] and thereafter, research in Survey Sampling Theory has been grouped into the research areas of Randomisation (Design Based Inference), Model Based Inference, Survey Design, Variance Estimation, General Theory, Analytical Inference, Estimating Distribution Functions, Handling Problems of Missing Data and Measurement of Errors.

Finite Population Framework

In this, the idea has been to assume that there exists a population frame, say U, consisting of N identifiable units, and a sample, say s, of n identifiable units is then defined as a subset of this U according to a rule

$$(1) \qquad p(s) = \begin{cases} \frac{1}{N_{C_n}}, & for\, all\, samples\, of\, size\, n. \end{cases}$$

See Chao [6] and Castledine [5].

Generally, to use a design other than simple random sampling, one requires auxiliary information, say;

$$X_i, i = 1, 2, 3, \ldots, N$$

for every unit in U.

Problems of this nature embraces stratification and clustering of units, techniques analogous to to blocking and nesting in experimental designs.

Going into further detail, it is worth mentioning that there could be, in addition to design variables, other auxiliary variables usually referred to as covariates, measured only on the sample units but for which population totals are known.

Remark.

Such covariates are usually employed to improve the efficiency of estimation.

Approaches to Sample Survey Inference

1. **Randomisation Inference**

   The framework of this inference is the distribution of the results of all the possible samples that could have been drawn using the random sampling scheme defined in equation 1.

   This distribution depends on the population matrix of the values of the survey variable which in general, is unknown.

   Control over the selection of units raises the question of whether or not some sampling schemes are better than others.

   For instance, Neyman (1934), laid down the foundations of randomisation inference and and addressed the issue of efficiency by considering optimal allocation for stratification problems.

In this case, it was assumed that the auxiliary information available was sufficient to group the elements into strata, and that the units within the strata were were homogeneous to the extent of allowing for simple random sampling within the strata.

The main results of Armitage [2] were that despite the analysis of variance partition of the total sum of squares into within and between components, the introduction of finite-population corrections means that for some populations: V_{SRS} is smaller than V_{PROP} and can even be smaller than V_{OPT}.

However, the general situation is that if the stratum sizes, say N_h are large, then

$$V_{SRS} \geqslant V_{PROP} \geqslant V_{OPT}.$$

2. **Model Based Inference**

Model Based Inference is due to Royall (1970). He argued that models can be used to make descriptive inferences on finite populations.

To do this, the population total, T, is written as

$$T = \sum_{s} y_i + \sum_{s^c} y_i \tag{2}$$

The idea is to use models to predict the unobserved values $y_i, i \notin s$ which is represented by the second part in equation 2, i.e. $\sum_{s^c} y_i$.

It has been shown that Model Based inferences can be much more efficient than the $p - based$ inferences. The argument for conditioning inference on the sample labels challenges the very basis of randomisation inference. For further readings on the works of Royall, see Royall and Eberhardt [9] and Royall and Cumberland [8].

3. **Model Assisted Inference**

To reconcile the $p-based$ and Model Based approaches, a generalised regression estimator

(3)

$$\hat{T}_{GREG} = \sum_s y_j/\pi_j + (\sum_u x_j - \sum_s x_j/\pi_j)\hat{\beta}^T$$

where $\hat{\beta}^T$ represents a row vector of estimated regression coefficients.

If one writes

$$e_{sj} = y_j - x_j\hat{\beta}^T \quad (4)$$

then the estimator defined in equation 3 can be written as

$$\hat{T}_{GREG} = \sum_U x_j\hat{\beta}^T + \sum_s e_{sj}/\pi_j. \quad (5)$$

Equation 5 is now a regression predictor plus and estimator of the total for the regression residuals. The estimator in equation 5 has been proved to have good properties in both the model based and $p-based$

approaches. This particular estimator can also be calibrated on the covariates.

In fact, presently, there is a choice between two classes of of variance for measuring the precision of estimation, the Randomisation Variance and the Model Based Variance.

4. **Design Assisted/Randomisation Assisted**

Sampling Theory to date

Sample Surveys are generated by several different processes. The finite population values may be generated by a model, which is always unknown, the sample is selected by some sampling mechanism, which in general is as defined in equation 1, which may be known, partially known or unknown.

The respondents are generated by another sampling process which is unknown and the measurements may be subject to measurement errors, also of unknown form.

The problem for the Statistician is usually to model this entire process and to make inferences that take into account all the sources of uncertainty, including those of the selection of models. Most significant results have been in the areas of Design of the variable probability, that is πps sampling schemes.

In the related problem of variance estimation, Quenouille's Jackknife has been quite influential.

Designs for balanced repeated replications variance estimators have also featured strongly and have linked work in surveys to related work in experimental design.

A similar link has been created in the area of optimal survey design. Model Based inference when employed as an alternative to the Randomisation approach to sampling, has and still places sample surveys inference within mainstream statistical inference.

The implications however, of the use of models for both design and analysis have challenged the basis of Randomisation Theory.

Purposive Designs have generally not seen a lot of success. The desire for procedures to be robust against model misspecifications supports the case for random designs and nonparametric estimators, and in this area, even empirical likelihood methods have emerged. In Analytic Inference from complex survey data, the key issue has been the role of selection, and hence of sample design.

It has been established that analyses based on simple random sampling assumptions, which ignore population structure such as clusters strata which are reflected in sample design, will usually be wrong.

Variances will be incorrect and tests and confidence intervals will be invalid.

Stratification on the dependent variable, as in response selective sampling, will almost always lead to estimation bias as well as to incorrect variances.

On missing values and their estimation, surveys are subject to nonresponse and this is an additional uncontrolled selection mechanism which is unlikely to be ignorable.

Imputation offers a possible solution. The conditions under which selection mechanisms of all types can be ignored for different forms of inference have also been clarified.

Techniques for estimation of distribution functions have also been advanced.

The work that has so far been done on analytic inference from complex data has placed sample survey theory within the area of mainstream statistical analysis, and this integration is likely to continue into the future.

Surveyors will adopt model based procedures, such as those now being employed for estimation for small areas of sub populations, while mainstream statisticians will take into account selection mechanisms when analysing observational studies.

Random designs will remain the preferred designs for most surveyors but conventional randomisation inference will be restricted to the estimation of finite population parameters from large-scale descriptive surveys. Notably, not much has been done regarding sampling in biological and environmental sciences.

The theory of capture-recapture methods is an exception, but this is not really a sample survey problem since the sampling cannot be controlled.

Particularly, Karl Pearson's aim of making the evolutionist a registrar-general for all forms of life, has not been achieved.

Current Debate Regarding Inference in Survey Sampling

If there exists choices for inference, that is prediction theory, probability sampling or even a hybrid of the two, a sample surveyor may have to make a decision on which one to use.

There is no doubt of the mathematical validity of either of the two approaches.

The Prediction Model has in most instances been referred to as a working model to emphasize that it is tentatively appropriate to be used.

Control and knowledge of probability sampling is complete (at least in principle) and therefore it is easy to see the appeal of basing inferences on it independent of prediction models.

The inevitable fallibility of our models has been an important theme of the theory of prediction based sampling.

There are however, a lot of fundamental issues underlying the background of the model based versus design based controversy.

For review articles that compare the theories, see Hansen et al. [7], Royall and Cumberland [8], Smith [10], Smith [11], and Smith [12].

A basic idea governing a great deal of statistical inference if the Conditionality Principle, in which we condition on observed random variables whose probability distribution is known and thus, not dependent on parameters about which we must make inferences.

The unconditional variance estimator is probabilistically correct but inferentially wrong.

It is inferentially wrong as a tool for helping to interpret and communicate the uncertainty in our estimate of the population mean.

The conditionality principle says that inference should be made conditionally on the observed sample, and not averaged over all samples that might have been selected, as the probability sampling approach does.

The Randomisation Principle says that random sampling is the **sine quo non** of finite population inference.

If we choose to reject that principle, then the begging question is why we use random sampling.

As is expected, there are several arguments for the use of randomisation, some are correct whereas some are not.

One extreme is that Artificial Randomisation provides the only basis for rigorous probabilistic inference and that in the absence of randomisation, valid probabilistic inferences are impossible.

However, it has been shown that the probability distribution determined by artificial randomisation is not appropriate even when it is available.

Therefore, to claim that, in general, probabilistic inferences are not valid when the randomisation distribution is not available is simply wrong.

RESEARCH PROSPECTS

Estimation in the presence of outliers

For a particular target population, let a working model hold for most of the population but a small percentage of units be contaminated by following a model whose mean or variance is far removed from that of the core model. Such units are known as outliers.

Variance estimation for the outlier resistant estimators of totals is an open area for research. Some researchers have done some work, for instance Lee (1991) discussed the method-of-moments types of estimators of the asymptotic variance and gave some limited empirical results.

Resampling has also been used as a possible way of addressing outliers.

For example, Chambers and Kokic (1993) used a version of bootstrap to calculate the confidence intervals in two populations, using the bisqure Ψ-function and three different Huber functions based on different values of the tuning constant.

Confidence intervals based on biased estimators have however been proved to to be centered in the wrong spot and will not, typically, have nominal coverage probabilities.

Even though the outlier robust alternatives may sometimes have good mean square error performance, there remains the problem of how to adequately correct for the bias when constructing confidence intervals.

Outliers will also affect the more general regression estimators. Substantial amount of research has been done in this area in other branches of statistics but direct links to finite population sampling has not been established.

Nonlinear Models

Today, surveys are mainly used to estimate rare characteristics, like the prevalence in a population of the number of persons who have a particular type of disability.

Such rare characteristics are ones for which linear models and linear estimators may be especially poor.

Little has been done on how much improvement can be made by using nonlinear estimators or whether it is feasible to use the estimators. Robust variance estimation for this class of problems has a lot of unfinished work. In particular, the work of Valliant (1986) has not been concluded. More specifically, no work has been done on testing whether balancing of some type has a role in robustness for Bernoulli models or other nonlinear models.

The Biotatistical literature on generalised estimation equations and accompanying variance estimators may be useful if properly applied to finite population estimation.

See the recommendations of Liang and Zeger (1986), Liang et al (1992), Zeger and Liang (1986).

Nonparametric Estimation of Totals

No research appears to have been done on the estimation of the variance of any of the nonparametric CDF estimators or on confidence interval constriction. How to use Multiple Explanatory Variables in predicting the population total and thus the corresponding CDF, say, $F_N(t)$, is a largely unexplored area.

Little has also been done on quantile estimation via nonparametric CDF estimators. Replication of some sort, like the bootstrap or balanced repeated replication, may be a possibility for variance estimation and confidence interval construction, particularly since analytic derivation of variances for nonparametric CDF and quantile estimators may be difficult.

In this area, an interested reader may find some insight from Shao and Wu (1992), Dorfman and Valliant (1993), Rao and Shao (1993). These works need to be extended to quantiles derived from parametric and nonparametric regression methods.

Small Area Estimation

A lot of literature in small area estimation exist, but use of such literature in finite population estimation has not been explored.

The unease in using concepts of small area estimation are mainly due to misuse of models. There is a lot of urge to guard inference against model failure by use of either model robust methods such as balanced sampling, or by careful and adequate model verification in the sampling context. Very little has been done in this area.

Verification is a very important aspect in small area estimation. An important component of this is cross-validation on small domains for which data outside of the domain and comparing predicted results on the domain with actual sample values.

The degree and type of cross-validation will depend on the amount and sort of data available within the domain.

In the case where a domain totally lacks data, one needs to investigate the validity of a priori justifications for applying the model, even when it is well verified on available data.

Such justifications need to be published along with any estimates. Much work needs to be done for developing a generally accepted canon of model verification and sound variance estimation for small area estimation.

References

Anthony, Y. C. K. (1993). A kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika*, 80(2):385–392.

Armitage, P. (1947). A comparison of stratified with unrestricted random sampling from a finite population. *Biometirka*, 34:273–280.

Bashtannyk, D. and Hyndman, R. (2001). Bandwidth selection for kernel conditional density estimation. *Computational Statistics and Data Analysis*, 36:279–298.

Beaumont, J. F. and Bocci, C. (2006). A practical bootstrap method for testing hypothesis from survey data. Technical report, Ottawa, Canada.

Castledine, B. (1981). A bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 68:197–210.

Chao, M. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69:653–656.

Hansen, M., Madow, W., and Tepping, B. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys.

Journal of the American Statistical Association, 73:776–793.

Royall, R. M. and Cumberland, W. G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76:66–82.

Royall, R. M. and Eberhardt, K. R. (1975). Variance estimates for the ratio estimator. *Sankhya C*, 37:43–52.

Smith, T. (1976). The foundations of survey sampling a review. *Journal of the Royal Statistical Society, Series A*, 139:183–198.

Smith, T. (1984). Present position and potential developments. personal views. sample surveys. *Journal of the Royal Statistical Society, Series A, Part 2*, 147:208– 221.

Smith, T. (1994). Sample surveys 1975, 1990. an age reconciliation. *International Review*, 62:3–34.