Selection always matters

T.M.Fred Smith[1] and Roger A. Sugden[2]

[1]*Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK. (tmfs@soton.ac.uk)*

[2]*School of Mathematical Sciences Queen Mary, University of London, London, E1 4NS, UK.*

Abstract

In 1977 Tim Holt and Fred Smith were funded by the UK Economic and Social Research Council for a programme of research into the Analysis of Complex Surveys. Gad Nathan was one of the visiting researchers. After reviewing the position of survey inference in the 1970's we consider some of Gad's work on regression analysis of survey data. An early belief was that random sampling selection could be ignored for model-based inference. This is explored and shown to be false under many circumstances. Our basic premise, following Rubin, is that all data models should include indicators for sample selection and for response. Both indicators should be modelled by the analyst, together with the survey variables, and only ignored if this follows from the models, the targets for inference and the type of inference; randomisation, frequentist, likelihood or Bayesian

Keywords: Inference from sample surveys, model-based inference, ignorable and informative designs, regression analysis, non-response.

1. Introduction

In 1977 Tim Holt and Fred Smith were funded by the UK Economic and Social Research Council for a programme of research into the Analysis of Complex Surveys. A key component of our proposal was that support should be provided for extended visits to Southampton from leading researchers in the area of complex survey analysis. One of our first invitations was to Gad Nathan who combined the skills of both theoretical research at the Hebrew University of Jerusalem and practical experience from the Israel Central Bureau of Statistics. Gad agreed to be one of our visiting researchers, and thus started a long collaboration which also included visits by Danny Pfeffermann, whose Ph.D under Gad's supervision had included the regression analysis of multi-stage survey data, extending the results in Scott and Smith (1969). In March 1986 the key results were presented in a 3-day conference at Southampton and the proceedings were published in the book 'Analysis of Complex Surveys' (Eds Skinner, Holt and Smith (1989)), including contributions by Gad.

2. Survey analysis in the 1970's

It is interesting to look back at survey analysis in the early 1970's. The work on foundations of finite population randomization inference by Godambe (1966), Basu (1971) and others had shown that there could be no optimal linear estimators and that for any random sampling design, $p(s)$, the likelihood function for the finite population vector $Y_U = (Y_1, ..., Y_N)$ of unknown fixed values was the constant, $p(s)$, for any population with the observed sample values $Y_i = y_i$ for i∈ s, and zero otherwise, and so was completely uninformative about the unobserved values Y_i for $i \notin s$. Instead of

trying to use the uninformative likelihood for inference by modelling the population data and conditioning on *s,* p-based inference averages over all the samples that might have been drawn employing concepts like approximate p-unbiasedness. For satisfactory inference it has to be assumed that the population values are constrained in some way so that the Central Limit Theorem will work. Essentially extreme values must be ruled out. But how can this be done without assuming a model for the population values? One reaction to these rather negative results was to propose model-based approaches to survey inference treating the population values as realizations of random variables generated by a parametric statistical model or super-population. Sample survey inference then conditions on the labels of the selected units and becomes a branch of mainstream inference. Exploring model-based approaches to survey inference was a major theme of the Southampton research programme. The targets for inference could be the super-population model parameters for analytic inference or finite population estimands, functions of finite population totals, for descriptive inference.

Samples can be selected in many ways. We consider only scientifically acceptable schemes such as office-based sampling schemes, carried out prior to data collection, which can be written as $p(s), s \in S,$ a set of known probabilities. In the early work on model-based approaches it was assumed that for sampling schemes determined before the sample was drawn the random variable, *s*, was independent of the target population values, $Y_U = (Y_1, ..., Y_N)$, and could be ignored for model-based inference. Royall (1970) considered sampling strategies, and showed that purposive sampling schemes would often be optimal and would still be ignorable for predictive inference of population totals.

The idea that random sampling schemes are ignorable, and the suggestion that purposive schemes might be used, was anathema to most survey statisticians. Wasn't it obvious that some sampling schemes were better than others and so must be related to the target population Y_U? Also random sampling was essential to avoid selection biases as well as for the measurement of sampling variation. In fact almost all those in favour of model-based approaches to survey analysis also argue in favour of random sampling for design. For if a sample is to be used by a large number of people then they must all be satisfied with the method for the selection of the units. Random selection is likely to be the only design that is acceptable to everybody, even if it is suboptimal for some. Basu (1971), a committed Bayesian, makes the case for randomization in terms of the control of unknown biases and the fact of the complexity of most surveys involving thousands of variables. Later Basu uses an argument at the base of scientific methodology: "I have no objection to prerandomization as such. Indeed, I think that the scientist ought to prerandomize and have the physical act of randomization properly witnessed and notarized. In this crooked world how else can he avoid the charge of doctoring his own data?"

One of the problems that Gad worked on was the regression analysis of survey data. In the 1970's the only option for most survey data analysts was to use a package such as SPSS or BMDP, and the default option for regression analysis was ordinary least squares (OLS), regardless of the sample inclusion probabilities. A sophisticated user might employ weighted least squares, weighting by the inverse of the sample inclusion

probabilities, but it was known that this gave the wrong estimates of variance. Assuming multivariate normality Pearson (1903) investigated the relationship between parameters before and after selection. Assuming selection on a known covariate, z, and that $\left(y_i^T, z_i^T\right)$ $i = 1,...,N$, are IID outcomes of the random normal vector $\left(Y^T, Z^T\right)$, then if (μ, Σ) are the moments before selection and $\left(\mu^*, \Sigma^*\right)$ are the moments after selection, they are related as follows:

$$\mu_y = \mu_y^* + \beta_{yz}\left(\mu_z - \mu_z^*\right), \quad \Sigma_{yy} = \Sigma_{yy}^* + \beta_{yz}\left(\Sigma_{zz} - \Sigma_{zz}^*\right)\beta_{yz}^T$$

where $\beta_{yz} = \Sigma_{yz}^* \Sigma_{zz}^{*-1}$. The moments after selection may be estimated by the equally weighted sample moments, and assuming the design variables are known we can find estimates of the mean and variance matrix before selection, and hence estimates of the regression coefficients between *y*-variables. These were termed *Pearson adjusted estimators*. Lawley (1943) showed that the same results follow under linear model assumptions.

Given the IID assumption the finite population moment matrices differ from the super-population matrices by terms of $O_p\left(N^{-1/2}\right)$, and so could be the targets for inference. But finite population targets can be estimated by p-based methods, leading to alternative p-based estimators of covariance matrices. Nathan and Smith (1989) present the results of several simulation studies comparing the OLS, model-based and p-based (model-assisted) estimators. For SRS or proportional stratified designs there are few differences, but for unequal probability stratified designs OLS is biased and the p-based estimator is relatively inefficient. The model-based estimator does well for all designs for multivariate normal data. However, these results are not robust to departures from linearity or homoscedasticity. See Nathan and Holt (1980) and Pfeffermann and Nathan (1981). Apart from ruling out OLS as a general purpose method of carrying out regression analysis with survey data, the results were not as clear cut as we had anticipated. Taking into account aspects of model uncertainty there is no clear winner. The main issue remains the fundamental one of whether inferences should average over repeated samples, *s*, or should condition on the particular realisation, *s*. The arguments and results of Royall and his colleagues strongly support the case for conditional inference. For Bayesians there is no alternative; you condition on known values, including *s*, and treat all unknowns as random.

3. **Ignorability and informative designs**

Random sampling raises questions about the uninformative nature of random designs and the conditions for their ignorability in model-based analyses. Rubin (1976), in a fundamental paper on inference and missing data, identifies sample selection with intentionally created or controlled missingness, while nonresponse is an example of unintentional or uncontrolled missingness. For parametric model-based approaches to inference he defines the concepts of missing at random, observed at random and distinct parameters and shows that the conditions for ignoring sample selection are different for frequentist, likelihood and Bayesian inferences. Typically random designs satisfy all the conditions, and so can be ignored, *provided that all the details of the*

design are known to the analyst. Analysis then proceeds from the marginal distribution of the sample data obtained by integrating out the missing data from the population model, in other words by employing the face-value distribution. This is the position adopted by most model-based data analysts.

Scott (1977) introduces the idea of design variables, z_U , related to the survey variables, y_U , known to the surveyor and used in determining the design. The selection scheme is now written, $p(s \mid z_U)$. The inclusion probabilities $\pi_i = \pi_i(z_U)$ are now related to the survey variables via the super-population model $p(y_U \mid z_U, \theta)$. In general the population likelihood is

$$p(y_U \mid z_U, \theta)\, p(I \mid z_U, y_U; \gamma)\, p(z_U; \phi) = p(y_U \mid z_U, \theta)\, p(s \mid z_U)\, p(z_U; \phi), \quad (1)$$

where *assuming full response, I* is the $N \times 1$ vector of inclusion indicators, $I_i = 1$ when unit *i* is observed and zero otherwise, and is determined by *s*. If y_{obs} is the matrix of observed values, and $y_U = (y_{obs}, y_{mis})$ then when z_U and $p(s \mid z_U)$ are known, integrating out y_{mis}, the sample likelihood of (θ, ϕ) is proportional to the face-value likelihood $p(y_{obs} \mid z_U; \theta)\, p(z_U; \phi)$. Under these conditions the sampling scheme is ignorable for model-based inference. However, the analyst may differ from the surveyor and may not have the same information about the design variables. In this case the joint distribution of all the data will involve averaging the population likelihood over the unobserved values in z_U, and in general the design is no longer ignorable and becomes informative to the analyst. Sugden and Smith (1984) explore this further and give conditions for ignorability which depend on the target and approach to inference. There is still work to be done in this area. As Scott (1977) says the only uniformly ignorable sampling scheme is simple random sampling.

As an illustration consider the commonly assumed IID super-population model and for simplicity assume that the only design information available to the analyst is the sample inclusion probabilities $\pi_i, i \in s$. How should the inference be modified? The correct likelihood is obtained by integrating the population likelihood (1) over all z_U that give the observed values of these probabilities. This depends on the details within the random sampling scheme $p(s|z_U)$. The general expression (in the IID case) is $\int \Pi_{i \epsilon s} p(y_i|z_i;\theta) p(s|z_U) \ \Pi_{i \epsilon U} p(z_i;\phi) dz_U$, where the range of integration is the set $z_U : \pi_i(z_U) = \pi_i, \ i \epsilon s$. In general this is difficult to carry out. However, when $\pi_i, \ i \epsilon s$ is the only design information, then if the range of integration splits into a product over sampled and unsampled units, the integration is straightforward. This is the case under unconditional Poisson sampling (UPS) where $p(s \mid z_U) = \Pi_{i \in s} \pi(z_i) \Pi_{j \notin s} (1 - \pi(z_j))$ with $\pi(.)$ a known function. Now $\pi_i(z_U) = \pi(z_i) = \pi_i$, and the likelihood is $\Pi_{i \in s} p(y_i \mid \pi_i; \theta, \phi) \pi_i \, p(\pi_i; \phi) (1 - \pi_\phi)^{N-n}$, where $\pi_\phi = E[\pi(Z)] = \int \pi(z_i) p(z_i; \phi) dz_i$ is a known function of ϕ.

Pfeffermann and Sverchkov (see their review paper of 2009) have a different approach. They start with the IID population model $\Pi_{i \in U} p(y_i)$ and relate the sample *pdf*

$p(y_i \mid I_i = 1)$ to the population *pdf* $p(y_i)$ under informative sampling. Their likelihood based on the sample distribution is $\Pi_{i \in s} p(y_i \mid I_i = 1; \theta, \phi)$, and via Bayes Theorem they express components as conditional expectations with respect to this distribution which enables their estimation using the sample data. They also require that the sampling scheme be expressible as products, as in the UPS scheme. Their results differ from the likelihood results above, even when the likelihood is conditioned on *s. W*hether the differences can be reconciled is an open question. Eideh and Nathan (2009) extend Pfeffermann and Sverchkov's results to informative two-stage sampling.

The above analysis has assumed that there is full response. This will rarely be the case. Non-response is an example of an uncontrolled sampling scheme and thus requires that the analyst models the unknown selection mechanism. The data collection process follows a distinct sequence; it starts with sample selection utilising known design variables which determines, *s;* unit non-response is a type of self-selection from the sample, *s*, and at this stage there may be only limited covariate data available. The final stage of data collection is administering the questionnaire, and this leads to uncontrolled item non-response. The corresponding response indicators are: $R_i = 1$ given $I_i = 1$ and $R_{j|i} = 1$ given $R_i = 1$, zero otherwise. Reversing the data collection sequence we first estimate missing items utilising the observed data and any covariates. Our preference, given the complex patterns of missing item data, is for some form of hot deck imputation. Given the completed item responses the next stage is to estimate the complete values y_i when $R_i = 0$, for the unit non-respondents, possibly by using post-stratification. This reconstructs the intended complete sample response, y_s. Then using multiple imputation the uncertainty due to estimating missing values can be captured. This reconstruction process, and the imputations, should be carried out by the sampler, who will not be subject to the constraints of confidentiality and will have access to a wide variety of auxiliary data. Given the completed sample(s) analysis can proceed using whatever inferential framework the analyst feels is appropriate.

In his discussion of Sarndal (2011), Brick (2011) states that "I believe that ignoring data collection was a major shortcoming of the debate that followed the work of Royall (1970) and his colleagues concerning the appropriate method of inference in sample surveys." This is unfair on Rubin, Little and their colleagues, who have done such fundamental work on missing values, including the idea of multiple imputation. As shown above reconstructing the complete sample response, y_s, means that at the next stage the inference is from y_s to y_U. There is still no agreement about how this should be done and a choice has to be made, not only between randomisation and model-based methods but also between alternative model-based approaches.

Conclusions

Control over selection is what distinguishes sample surveys from observational studies, and hence selection always matters. In this Year of Statistics we should admit the limitations of some of our methods, in particular the failure of the results of so many observational studies to replicate and the fact that large data does not necessarily mean more accurate data,; it will depend on how the data are selected and collected. Within

selection schemes the scientific case for random sampling is overwhelming. However, ignorability of selection is more complex, and just as randomization should be witnessed and notarized, so ignorability should be established in each case by listing the information available to the analyst and demonstrating that the conditions for ignorability are satisfied for the chosen framework for inference. If the sampling scheme is not ignorable and model-based analyses are employed then where possible these should use the full likelihood. Different designs do give different likelihoods for the same sample data, the calculations being very complicated for many designs.

References

Basu, D. (1971). An essay on the logical foundations of survey sampling, part one. *Foundations of Statistical Inference,* Holt, Reinhart and Winston, Toronto, 203-242.

Brick, J. M. (2011). Discussion of Sarndal (2011). *Journal of Official Statistics,* **27,** 23-28.

Eideh, A. and Nathan, G. (2009). Two-stage informative cluster sampling –estimation and prediction with applications for small area models. *Journal of Statistical Planning and Inference,***139**, 3088-3101.

Godambe, V.P. (1966).A new approach to sampling from finite populations. I. Sufficiency and linear estimation. *Journal of the Royal Statistical Society, B*, **28**, 310-328.

Lawley, D.N. (1943). A note on Karl Pearson's selection formula. *Proceedings of the Royal Society of Edinburgh,* **A62,** 28-30.

Nathan, G. and Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society,* **B42,** 377-386.

Nathan, G. and Smith, T.M.F. (1989). The effect of selection on regression analysis. *Analysis of Complex Surveys (eds Skinner, Holt and Smith).*Pps 149-163.

Pearson, K. (1903). On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society,* **A200,** 1-66**.**

Pfeffermann, D. and Nathan, G. (1981). Regression analysis of data from complex samples.*Journal of the American Statistical Association,* **76,** 681-689.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. Ch 39 in *Handbook of Statistics: Sample Surveys: Inference and Analysis*, Vol 29B, 455-487.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika,* **57**, 377-388.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581-592.

Sarndal, C-E. (2011). Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics,* **27,** 1-21. (The 2010 Morris Hansen Lecture).

Scott,A.J. (1977). Some comments on the problem of randomisation in survey sampling. *Sankhya,* C, **39**, 1-9.

Scott, A.J. and Smith, T.M.F. (1969). Estimation in multi-stage samples. *Journal of the American Statistical Association*, **64**, 830-840.

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

Sugden, R.A. and Smith, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, **71**, 495-506.