Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Survey Sampling In Official Statistics - Some Thoughts On Directions

Ray Chambers[1]

Abstract

I present a modeller's perspective on the current status quo in official statistics surveys-based inference. In doing so, I try to identify the strengths and weaknesses of the design and model-based inferential positions that survey sampling, at least as far as the official statistics world is concerned, finds itself at present. I close with an example from adaptive survey design that illustrates why taking a model-based perspective (either frequentist or Bayesian) represents the best way for official statistics to avoid the debilitating 'inferential schizophrenia' that seems inevitable if current methodologies are applied to the emerging information requirements of today's world (and possibly even tomorrow's).

Key Words: Design-based inference; Model-assisted inference; Model-based inference; Calibrated weighting; Calibrated Bayes; Adaptive design.

1. Introduction

This paper presents an overview of different approaches to official survey sampling (OSS) inference that are currently in use in official statistics, focussing on their relevance to ongoing changes to the environment for official statistics currently based on survey sampling methods. In particular, I will describe my perspectives on OSS Past, or design-based inference; OSS Present, or model-assisted inference; and OSS Future, or model-based inference. In doing so, I will emphasise the paradigm shift in official statistics inference that is now taking place (from design-based and model-assisted to model-based), and the implications for official statistics methodology looking forward.

As an example of this shift I discuss below how model-based thinking is useful when developing a strategy for adaptive survey design for informative non-response. However, this is but one of many such examples. In my case I have used model-based ideas in small area estimation, analysis of probability-linked data, population inference using new sources of auxiliary information, e.g. information on population networks, as well as when information from sample surveys and from population registers need to be combined.

My motivation for considering these issues is in large part a consequence of comments by Frauke Kreuter in her presentation at the 2013 Graybill Conference, where she noted " ... in recent years large survey organizations have made considerable efforts to enhance information on all sample cases with paradata, data from commercial vendors, and through linkage to administrative data to allow for improved field operations or non-response adjustments." In my subsequent President's letter in the Newsletter of the International Association of Survey Statisticians in June 2013, I built on these ideas, noting that "Sampling inference will have to adapt to this new data collection paradigm, with the importance of sampling error much diminished, and a real need to come to grips with how basic ideas like uncertainty should be characterised in the resulting mix of non-response errors, linking errors, measurement errors and model specification errors."

This paper is an attempt to more clearly delineate what I meant by these comments. In order to do so, however, it is useful to first develop some basic concepts and notation.

[1]National Institute for Applied Statistics Research Australia, University of Wollongong, Australia, 2522 (ray@uow.edu.au)

2. Basic Concepts

The concept of a finite population lies at the core of official statistics data collection. This is the population of N units for which information is required. This in turn implies the existence (at least in concept) of a list U of the N units making up this population, such that each population unit is identifiable on this list in the sense that the list contains a unique identifier or label for each population unit. Without loss of generality we index the finite population by this identifier, so it takes the values $i = 1,\ldots,N$.

Together with this list, we can, without loss of generality, also assume that we know the population values Z_i of an auxiliary variable Z. This may be nothing more than a population unit's identifier. However, in most cases it will be considerably more. The most important thing though is that each population value of Z can be associated with a unique value Z_i of the population identifier. Without loss of generality, we assume that Z is scalar. Finally, we shall characterise the unknown population values of interest by the values Y_i of a scalar target variable Y. The primary objective then is to infer the value of a well-defined function Q of the population values of Y and Z. In this context, the population total of Y is often taken as this objective, but this is only one of many potential objectives for an official statistics data collection exercise.

A fundamental scientific methodology used in OSS is random sampling. That is, for each unit i in the finite population, the survey sampler generates a value I_i for a random indicator variable I which equals 1 (0) if that particular population unit (i.e. label) is sampled (not sampled). Let $s = \{i; I_i = 1\}$ denote the labels of the n sampled population units. We can then define the population vectors $\mathbf{I}_U$, the N-vector of population values of I; $\mathbf{Z}_U$, the N-vector of population values of Z; $\mathbf{Y}_U$, the N-vector of population values of Y; and $\mathbf{Y}_s$, the n-vector of sample values of Y. Note that the distribution of $\mathbf{I}_U | \mathbf{Z}_U$ defines what is generally referred to as the design of the sample survey. See Smith (1983), Sugden and Smith (1984) and Pfeffermann (1993).

Random sampling is fundamental to OSS because it guarantees non-informative sampling given $\mathbf{Z}_U$. This is any method of sampling such that the distribution of $\mathbf{I}_U$ depends only on the known population values in $\mathbf{Z}_U$. That is, the distribution of $\mathbf{I}_U$ is completely specified given $\mathbf{Z}_U$. It immediately follows that under non-informative sampling given $\mathbf{Z}_U$ the distributions of $\mathbf{I}_U$ and $\mathbf{Y}_U$ are conditionally independent given $\mathbf{Z}_U$.

We note that the assumption of non-informative sampling is not appropriate where there is the possibility of selection bias, and consequent dependence between the distributions of $\mathbf{Y}_U$ and $\mathbf{I}_U$, even after conditioning on $\mathbf{Z}_U$. Selection bias almost always only of concern when the sampling is carried out by 'someone else' (the secondary analysis situation) or when the selected sample and the realised sample are not the same, due primarily to non-contacts and non-response. It is this second situation that is typically of most concern to OSS.

3. The Role of Design-Based Inference in OSS

Design-based inference is currently the standard inferential paradigm adopted in official statistics, and represents the underlying philosophy of most standard texts on survey sampling theory. Its origin is usually associated with Neyman (1934), and it treats all population values as generated by 'nature', i.e. $\mathbf{Y}_U$ and $\mathbf{Z}_U$ are vectors of finite population parameters and are therefore conditioned on in any inference. Any unknown function of population values (e.g. Q) is then also a finite population parameter whose value has to be inferred from the sample data. In particular, it is assumed that there is a function $\hat{Q}(\mathbf{Z}_U, \mathbf{Y}_s)$ of these data that is an estimator of Q.

Given this set up, it is clear that the only random variable that can be used for inference about Q given $\hat{Q}$ is the one whose distribution is (at least in theory) completely known, i.e. $\mathbf{I}_U$. In effect, inference is with respect to the potential values of $\hat{Q}(\mathbf{Z}_U, \mathbf{Y}_s)$ that can be generated given the possible values of $\mathbf{I}_U$ (and hence s) and the fixed values of $\mathbf{Y}_U$

and $\mathbf{Z}_U$. Valid inference requires design unbiasedness (if possible) or design consistency (at least), in the sense that for any choice of a distribution for $\mathbf{I}_U$ (the sampling process), the distribution of the random variable $\hat{Q}$ should be such that $E\left(\hat{Q}-Q\middle|\mathbf{Z}_U,\mathbf{Y}_U\right)\approx 0$. Efficient inference then requires one to identify a distribution for $\mathbf{I}_U$ that makes $E\left(\left(\hat{Q}-Q\right)^2\middle|\mathbf{Z}_U,\mathbf{Y}_U\right)$ as small as possible, typically subject to a cost constraint. Where no restriction is placed on $\mathbf{Y}_U$, this is an impossible task, as the non-existence result of Godambe (1955) makes clear. See Basu (1971) for an accessible proof.

As noted earlier, design-based survey sampling theory tends to focus on estimation of the population total T of Y, and in particular on linear estimation of this total, so that the preferred estimator of T has the form

$$\hat{T}=\sum_{i\in s}w_iY_i=\sum_{i=1}^{N}w_iI_iY_i\ .$$

Here w_i is a weight that depends only on the auxiliary data $\mathbf{Z}_U$. Since the distributions of w_i and I_i are completely determined by $\mathbf{Z}_U$, the necessary condition for design unbiasedness is

$$E(w_iI_i\,|\mathbf{Z}_U,\mathbf{Y}_U)=E(w_iI_i\,|\mathbf{Z}_U)=1$$

or the more familiar Horvitz-Thompson definition $w_i=\pi_i^{-1}$, where $\pi_i=E(I_i\,|\mathbf{Z}_U)$. That is, even though this choice of weight may not be efficient, it will always lead to valid inference. As a consequence, the Horvitz-Thompson estimator, and its variants, have been fundamental to design-based OSS since the 1940s.

A reality check is appropriate at this stage. Inadequate sampling frames, non-response and contact/response targeting by both interviewers and survey managers all mean that the actual distribution of $\mathbf{I}_U$ is almost never equal to its 'design' distribution. This has been known by statisticians working in OSS from the beginning, and constitutes a huge challenge for proper OSS implementation of design-based inference. Deviation of the distribution of the indicators for those population units actually providing survey data from the distribution of sample inclusion indicators determined by the sample design implies that the sampler does not have full control of the sampling mechanism underpinning the realised (rather than selected) sample, and so does not 'know' its properties. In particular, the crucial non-informative sampling assumption $E(I_i\,|\mathbf{Z}_U,\mathbf{Y}_U)=E(I_i\,|\mathbf{Z}_U)$ may then no longer be valid. In this situation valid inference based on Horvitz-Thompson weighting may no longer be possible. Another approach to inference is required.

4. The Role of Model-Assisted Inference in OSS

As the proof of Godambe's non-existence result makes clear, the main problem with design-based inference is that it is too general, since its properties have to hold for any $\mathbf{Y}_U$. OSS practitioners have from the start, and certainly by the 1940s, therefore adopted sampling strategies that are, in some sense, reasonable over the space of potential realisations of $\mathbf{Y}_U$. Such values of $\mathbf{Y}_U$ are typically identified by assuming a model for the distribution of $\mathbf{Y}_U$ given $\mathbf{Z}_U$. This is used in specification of the sampling scheme (definition of strata, clusters, inclusion probabilities) as well as in specification of the estimator (ratio, regression estimators).

With the additional assumption of a model for $\mathbf{Y}_U$ given $\mathbf{Z}_U$, it is theoretically possible to identify optimal sampling designs by minimising the model expected value of the design mean squared error (MSE). However, this is rarely done. Instead, the model is used the create calibrated weights for model-assisted survey estimation, where the 'design' weights $w_i=\pi_i^{-1}$ are modified to recover known population counts (benchmarks) from the observed sample data. Since choice of which population benchmarks to recover is equivalent to specifying a linear model for $\mathbf{Y}_U$ in terms of the variables defining the benchmarks, population modelling necessarily guides this choice. A consequence is that the estimator of T based on these calibrated weights is then an unbiased estimator of the expected value of T under this model.

However, inference remains design-based under the model-assisted approach. In particular, the key requirement of validity is still in force, so design unbiasedness and consistency requirements are viewed as crucial. That is, although models are used to identify estimation strategies, any consequent inference remains design-based, at least in principle.

This perceived validity is often used to justify the claim that model-assisted OSS inference is robust to model misspecification because its requirement of exact or approximate design-unbiasedness ensures correct inference irrespective of the 'true' nature of the distribution of $\mathbf{Y}_U$ given $\mathbf{Z}_U$. To borrow the words of a friend and colleague (Ken Brewer), "Adopting the model-assisted approach is like wearing both a belt and braces to hold up one's trousers. If the belt (the model) should break, then one is not going to be totally embarrassed, since the braces (design-unbiasedness) should still keep things in place."

This is reasonable, since any model specification is wrong. However, it ignores the fact that the resulting inference can be very inefficient. A slight adaptation of Kendall (1959) seems to me to capture the essence here. Following Kendall's witty revision of Longfellow's epic poem, we observe that Hiawatha puts his faith in randomisation rather than shooting practice when attempting to win an archery contest, on the basis that his shooting technique is expertly randomised, and his arrows are all model-assisted, so his archery strategy is design consistent and is consequently very close to being unbiased for the true target. When the inevitable happens, and he comes last in the contest, Hiawatha retreats into the forest, where

"In a corner of the forest
Sits alone my Hiawatha
Permanently cogitating
On the normal law of errors.
Wondering in idle moments
If perhaps increased precision
Might perhaps be sometimes better
Even at the cost of bias,
If one could thereby now and then
Register upon a target."

5. The Role of Model-Based Inference in OSS

Unlike the design-based approach, the model-based approach in OSS has no obvious starting point. Brewer (1963) is perhaps the earliest attempt to explicitly base inference on a model in a survey sampling context. However, by far the most influential work on this approach was independently carried out by Royall and his colleagues from about 1969. See Royall (1976a) for a clear and thorough development of the basic ideas, underpinned by the observation that the non-informative sampling assumption means that $\mathbf{I}_U$ is ancillary for inference about $\mathbf{Y}_U$ given $\mathbf{Z}_U$, and so inference about Q should therefore be conditional on $\mathbf{I}_U$ and $\mathbf{Z}_U$, and not $\mathbf{Y}_U$ and $\mathbf{Z}_U$.

Under the model-based approach, a natural requirement for a linear $\hat{Q}$ is model-unbiasedness. That is, $E\left(\hat{Q}-Q|\mathbf{I}_U,\mathbf{Z}_U\right)=0$. Furthermore, uncertainty is quantified by the mean squared error of prediction (MSEP), $E\left(\left(\hat{Q}-Q\right)^2|\mathbf{I}_U,\mathbf{Z}_U\right)$. Standard statistical prediction arguments allow one to then define optimal weights for use in sample estimation and inference. In effect, survey sampling inference becomes a subset of standard statistical prediction theory.

Model-based ideas have been largely ignored (or resisted) within OSS. In large part this appears to be because of the comforting observation that, given a linear model for $\mathbf{Y}_U$, the model-unbiasedness property for a linear estimator is equivalent to using calibrated weights, where the calibration is with respect to the population totals of the model covariates. So model-assisted weighting is viewed as being model-based 'in spirit'. However, it is also because in the early stages of the development of the model-based approach much was made of the fact that assuming a model for $\mathbf{Y}_U$ implied the existence of an optimal sample that minimised the MSEP of an estimator. This can be viewed as of no great consequence in many commonly used population models, e.g. when the regression relationship between Y

and Z is linear with a constant error variance, in which case the optimal sample for the regression estimator of T is a balanced one where the sample mean of Z equals the population mean of Z, a property that holds in expectation under simple random sampling. However, it is also conceptually unsettling, since it seems to remove the fundamental need for randomisation in survey sample-based inference. In particular, under the widely-used linear 'regression through the origin' model for Y in terms of a size variable Z, where the variance of Y given Z is proportional to a non-negative power of Z, the optimal sample for the optimal model-based estimator of T consists of the n population units with the largest values of Z. This was generally viewed as a step too far from an OSS perspective, though see Karmel and Jain (1987) for an explicit, and positive, OSS-based evaluation of this strategy.

Development of Bayesian approaches to model-based sample survey inference followed quickly once the prediction theory framework for this inference was established. Here we summarise this approach, where p is used to denote the population data distribution under the assumed model, and $prior$, $post$ are used to denote the prior and posterior distributions respectively for the unknown quantities in the model. With this specification, the posterior predictive distribution for $Q = Q(\mathbf{Y}_U, \mathbf{Z}_U)$ is then written as

$$post\left(Q|\mathbf{Y}_s, \mathbf{I}_U, \mathbf{Z}_U\right) \propto \int \underbrace{p\left(Q|\mathbf{Y}_s, \mathbf{I}_U, \mathbf{Z}_U; \theta, \phi\right)}_{\text{model-based conditional density of } Q} post\left(\theta, \phi|\mathbf{Y}_s, \mathbf{I}_U, \mathbf{Z}_U\right) d\theta d\phi .$$

Under non-informative sampling one can show that $p\left(Q|\mathbf{Y}_s, \mathbf{I}_U, \mathbf{Z}_U; \theta, \phi\right) = p\left(Q|\mathbf{Y}_s, \mathbf{Z}_U; \theta\right)$ and

$$post\left(\theta, \phi|\mathbf{Y}_s, \mathbf{I}_U, \mathbf{Z}_U\right) \propto \underbrace{p\left(\mathbf{Y}_s|\mathbf{Z}_U; \theta\right) prior\left(\theta|\mathbf{Z}_U\right)}_{post\left(\theta|\mathbf{Y}_s, \mathbf{Z}_U\right)} \underbrace{p\left(\mathbf{I}_U|\mathbf{Z}_U; \phi\right) prior\left(\phi|\mathbf{Z}_U\right)}_{post\left(\phi|\mathbf{I}_U, \mathbf{Z}_U\right)} .$$

Consequently

$$post\left(Q|\mathbf{Y}_s, \mathbf{I}_U, \mathbf{Z}_U\right) \propto \int p\left(Q|\mathbf{Y}_s, \mathbf{Z}_U; \theta\right) p\left(\mathbf{Y}_s|\mathbf{Z}_U; \theta\right) prior\left(\theta|\mathbf{Z}_U\right) d\theta .$$

In general this function cannot be analytically determined. However it can be computed via simulation, using widely available software (e.g. BUGS / BRUGS).

More recently, signs of an apparent reconciliation between the design-based but model-assisted and the 'full' model-based ideologies has started to appear, in the form of what Little (2006, 2012) refers to as Calibrated Bayes. Desiderata for this approach hinge around the requirement that any models used in Bayesian analysis of sample survey data should incorporate sample design features. In particular, the 'Calibrated Bayesian' should use robust models with good repeated sampling properties, since models that ignore features like survey weights are vulnerable to misspecification. The basic aim then is to guide Bayesian model specification towards models that incorporate design features, and so lead to inferences with good design-based properties. Some guidelines for doing this include using design weights as covariates in the prediction model (e.g. Gelman, 2007), and modelling multistage sample designs using multilevel random effects models, an idea that goes back to Royall (1976b) and is very much the 'industry standard' in small area estimation at present.

6. Stepping Back

It is useful at this stage to step back and take stock. It is an incontrovertible fact that probability sampling and the law of large numbers ensures that a design-consistent estimator of Q will have a very small probability of being far from to its true value if the sample is large. However, how large does 'large' have to be before asymptotics become relevant? I would hazard a guess that asymptotics are irrelevant for most OSS survey sampling outputs. Furthermore, the widespread use of model-based small area estimation raises the uncomfortable question - if model-based inference is good enough for small samples, why is it not good enough for large samples? This curious dichotomy of inferential 'respectability' within OSS has been labelled "inferential schizophrenia" by Little (2012). Finally, I note that the robustness of validity conferred by the use of design-based inference is not robustness of efficiency. In particular, I have often observed that methods with good large sample design-based properties can be (and often are) inefficient for any particular sample size.

But there is no free lunch. Model-based inference is not valid under model misspecification, so a good model specification search is vital. If the model is 'invalid' then the optimal sample design (and estimation) strategy based on it can lead to large errors. The important thing to note here is what is meant by choice of an invalid model. The oft

quoted statement (usually attributed to George Box) that "all models are wrong, but some are useful" applies. It is misleading to state that adoption of a model-based inferential approach requires one to somehow decide before one sees the sample data that one specific model specification for the relationship between $\mathbf{Y}_U$ and $\mathbf{Z}_U$ should be the basis for any further inference about Q. The results in Hansen, Madow and Tepping (1983) illustrate the consequences of adopting a sampling strategy that is not model-robust and a working model specification that ignores the information about model breakdown in the sample data. Misspecification robust modelling needs care. Which explains why careful use of stratification (pre- and post-) is so important, as is inclusion of relevant covariates. In this context the usual advice about the importance of model diagnostics, and the application of data adaptive modelling strategies (including stratification and multiple covariates, when available) is key.

One data adaptive modelling strategy is non-parametric modelling. That is, we use non-parametric methods to model $E\left(\mathbf{Y}_U \middle| \mathbf{Z}_U\right)$, assuming a flexible model specification of the form

$$\mathbf{Y}_U = m(\mathbf{Z}_U) + \mathbf{e}_U$$

where $\mathbf{e}_U$ is the population vector of model errors and m is a sufficiently smooth function of the covariates in $\mathbf{Z}_U$. Popular methods for fitting m include kernel-based methods and spline-based methods, both of which allow the final estimator to be written in linear (i.e. weighted) form.

A problem that rises immediately with this approach is that these non-parametric weights are not calibrated in general. This is easily dealt with, but it is instructive to compare the model-assisted and model-based approaches to this calibration.

The model-assisted calibration approach is usually associated with Deville and Särndal (1992). The idea here is to choose a vector of sample weights $\mathbf{w}_s^{cal}$ that is close to the vector of nonparametric sample weights $\mathbf{w}_s^{np}$ but at the same time satisfy the calibration constraints (i.e. weights $\mathbf{w}_s^{cal}$ are calibrated on the columns of $\mathbf{Z}_U$). This of course is equivalent to a parametric correction for the bias under the linear model specified by the calibration constraints. In any case, a commonly used metric for 'closeness' is the Euclidean metric

$$V = (\mathbf{w}_s^{cal} - \mathbf{w}_s^{np})^T \, diag(\mathbf{v}_s)(\mathbf{w}_s^{cal} - \mathbf{w}_s^{np})$$

where $\mathbf{v}_s$ is a specified set of constants that is supposed to characterise the heteroskedasticity in the residuals generated by this model. Minimising V subject to calibration leads to calibrated sample weights of the form

$$\mathbf{w}_s^{cal} = \mathbf{w}_s^{np} + \underbrace{\mathbf{H}^T(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_s^T\mathbf{w}_s^{np})}_{calibration\,correction}$$

where $\mathbf{H}$ is a matrix such that $\mathbf{H}\mathbf{y}_s$ defines the estimator of the vector of linear model coefficients under the model. Note that setting the nonparametric weights equal to Horvitz-Thompson type weights (i.e. the inverses of the sample inclusion probabilities) leads to the generalised regression (GREG) estimator. In any case, the calibrated estimator defined by $\mathbf{w}_s^{cal}$ will be model-unbiased under the linear model defined by $\mathbf{Z}_U$.

From a model-based perspective, calibration has a somewhat different interpretation. This approach starts with a working linear model defined by the columns of $\mathbf{Z}_U$. However, we recognise that this model is only an approximation to reality, and so prudently add a nonparametric bias correction to the optimal regression estimator under this (approximate) working model. Given this set up, Chambers, Dorfman and Wehrly (1993) suggest that the sample residuals be used to compute a nonparametric estimate of the bias, and then this bias estimate be subtracted from the original optimal model-based estimate. That is, given a working model of the form $\mathbf{y}_s = \mathbf{Z}_s\boldsymbol{\beta} + \mathbf{e}_s$, where $\mathbf{e}_s \sim \left(\mathbf{0}, diag(\mathbf{v}_s)\right)$, we use standard least squares to obtain fitted values $\hat{\mathbf{y}}_s = \mathbf{Z}_s\mathbf{H}\mathbf{y}_s$ and residuals $\mathbf{r}_s = (\mathbf{I}_n - \mathbf{Z}_s\mathbf{H})\mathbf{y}_s$. We then use the vector $\mathbf{m}_s$ of nonparametric weights to estimate the population total of these residuals, and add this estimate to the original linear model optimal regression estimator. This is of course equivalent to using the model-based non-parametrically bias corrected weights

$$\begin{aligned}\mathbf{w}_s^{p+np} &= \underbrace{\mathbf{1}_n + \mathbf{H}^T(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_s^T\mathbf{1}_n)}_{parametric\ (BLUP)\ weight} + \underbrace{(\mathbf{I}_n - \mathbf{H}^T\mathbf{Z}_s^T)\mathbf{m}_s}_{nonparametric\ bias\ correction} \\ &= \mathbf{1}_n + \mathbf{H}^T(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_s^T\mathbf{1}_n) + (\mathbf{I}_n - \mathbf{H}^T\mathbf{Z}_s^T)(\mathbf{w}_s^{np} - \mathbf{1}_n) \\ &= \mathbf{w}_s^{np} + \underbrace{\mathbf{H}^T(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_s^T\mathbf{w}_s^{np})}_{calibration\ correction}.\end{aligned}$$

That is, one ends up with exactly the same correction as the model-assisted calibrator. Since the efficacy of the model-based approach depends on the nonparametric bias correction, we also see that the set of weights that one chooses as the starting point for calibration matters considerably.

7. Implications for design and estimation in OSS

There is a very practical problem that must first be overcome before model-based ideas can hope to gain much traction in OSS. This relates to the fact that OSS surveys are typically multipurpose, collecting many different items of information measured on different scales, which implies many different models and associated optimal sample designs. It is impossible to impose these simultaneously. In this context, stratification, systematic selection (if population can be ordered), random selection (when it can't) are all useful tools for creating multipurpose designs that lead to samples that can claim to 'represent' a population.

As far as I know, no metric for representativeness exists, but a reasonable approximation is to choose a balanced sample, i.e. one where the distribution of values in $\mathbf{Z}_s$ echoes that in $\mathbf{Z}_U$. In most cases this means selecting a sample that is at least approximately calibrated to a working model based on $\mathbf{Z}_U$, i.e. one such that sample estimates of known population benchmarks are exact. However, I also note recent work on model-robust sampling designs by Welsh and Wiens (2013), where a sample design (i.e. a value for $\mathbf{Z}_s$) is chosen in order to minimise the maximum prediction error that could occur if the real model is in a neighbourhood of the working model.

At this stage it seems reasonable to question the role randomisation in sample design under the model-based approach. Clearly it is an integral part of both the design-based and the model-assisted approaches. Without randomised sample selection there is no inference in these cases. In contrast model-based inference places no restriction (beyond non-informativeness) on the actual method used to select the sample. What matters are the characteristics of the chosen sample (e.g. its balance characteristics). But, unless strictly controlled, probability sampling can also lead to samples that are very non-robust from a model-based perspective. Consequently use of probability sampling methods that allow sample characteristics to be controlled seems a good thing. In this context one can note that constrained random sampling has played an important role in design-based sampling, see Tam and Chan (1984) and Deville (1992) for a discussion of rejective sampling and Deville and Tillé (2004) for cube sampling, which allows selection of random samples that are also balanced.

However, the proper role of randomisation under the model-based approach is not to ensure exact balance (which it cannot). What it does guarantee, however, is non-informativeness, thus allowing valid model assessment from the sample data. That is, one can view the appropriate role of randomisation as not so much to ensure valid design-based inference, but actually to ensure valid model-based inference. In particular, randomisation ensures non-informative sampling when used to define the distribution $p(\mathbf{I}_U|\mathbf{Z}_U;\phi)$. Its protective characteristic is then that it balances on missing covariates in $p(\mathbf{Y}_U|\mathbf{Z}_U;\theta)$ 'on expectation'.

8. Operationalizing model-based inference within an OSS environment

By definition, model-based inference requires model specification, i.e. definition of $\mathbf{Z}_U$. In this context, the requirement for non-informative sampling, i.e. $\mathbf{Y}_U \perp \mathbf{I}_U|\mathbf{Z}_U$, is as much a statement about $\mathbf{Y}_U$ and $\mathbf{Z}_U$ as it is a statement about $\mathbf{I}_U$. In particular, it requires that we define $\mathbf{Z}_U$ so that it includes the covariates that determine $\mathbf{I}_U$. At a minimum, therefore, stratification and clustering should be 'built into' $\mathbf{Z}_U$. Survey weighting should also be built

into $\mathbf{Z}_U$ by including the covariates that determine these weights; or if these design covariates are unavailable, the weights themselves.

The drawback of this approach is a potentially inefficient prediction model, containing too many irrelevant predictors for any particular target variable. Here model diagnostics can be used to decide which aspects of the sample design to keep in the model. Furthermore, there is now an extensive literature on methods for fitting overspecified models (e.g. ridging, regularisation) which essentially aim to cut back on model 'degrees of freedom'. See Clark and Chambers (2008) for an exploration of model specification within a sample survey context.

At the end of the day, however, there will always still remain the issue of whether $\mathbf{Z}_U$ is 'correctly' specified. The main reason for this is non-response. How can the statistician be sure that the method of sampling, including the involuntary sampling due to non-response, is non-informative given the variables included in $\mathbf{Z}_U$? This issue is even more of a concern for secondary analysts, who often lack access to important design variables (beyond strata, cluster and weighting information). Consequently, it behoves official statisticians to ensure that any OSS data released for public modelling contain sufficient auxiliary information to ensure that they can be analysed as if they are the result of a non-informative sampling process

Which brings us to the take home message: Good sample design controls what needs to be controlled (in the sense of defining $\mathbf{Z}_U$ and $\mathbf{Z}_s$ to ensure decent prediction of T) and randomises over what cannot be controlled. Hopefully, what is controlled but doesn't need to be (the components of $\mathbf{Z}_U$ that can be dispensed with), has little impact on inference. Even more hopefully, the randomisation has ensured that any missing components of $\mathbf{Z}_U$ (those associated with non-contacts and non-response and essentially non-controllable) are effectively balanced across the realised sample, and so have little impact on prediction of T .

9. An example of model-based thinking in OSS

The question that we (briefly) address in this last section is simply put: How does one adapt the follow-up strategies used in a multi-wave contact + interview survey for potential non-ignorable non-response? In order to see what can be done here, we assume that the working model for the survey data is the standard linear model

$$\xi: \quad \mathbf{y}_U = \mathbf{Z}_U\boldsymbol{\beta} + \mathbf{e}_{\xi U}.$$

Here $\mathbf{Z}_U$ is a known matrix of population covariates, $\mathbf{e}_{\xi U} \sim (\mathbf{0}, \sigma_\xi^2 \mathbf{I}_N)$ and it is well known that the optimal sampling weights under ξ are $\mathbf{w}_{\xi s} = \mathbf{1}_n + \mathbf{Z}_s\left(\mathbf{Z}_s^T\mathbf{Z}_s\right)^{-1}\left(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_s^T\mathbf{1}_n\right)$. Use of ξ for survey estimation is justified provided sampling is non-informative given $\mathbf{Z}_U$, i.e. $\mathbf{e}_{\xi U} \perp \left(\mathbf{I}_U, \mathbf{Z}_U\right)$. This is hard to justify (at least theoretically) if there is non-response (non-contacts and/or refusals). With non-response follow-up, however, there is an opportunity to 'target' particular non-respondents in order to reduce non-response bias.

Suppose that there are m observed (i.e. responding) sample units and that their working model-based weights are $\mathbf{w}_{\xi o} = \mathbf{1}_m + \mathbf{Z}_o\left(\mathbf{Z}_o^T\mathbf{Z}_o\right)^{-1}\left(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_o^T\mathbf{1}_m\right)$. Here we use o to denote the observed sample. The problem that we then face is that these weights do not define an unbiased estimator of the population total T , since $\mathbf{e}_{\xi U} \perp \left(\mathbf{R}_U, \mathbf{I}_U, \mathbf{Z}_U\right)$ does not hold.

Suppose that this bias is due to missing covariates, and there exists a variable X , observable only on the sample, and with samples values $\mathbf{X}_s$, such that the 'true' model for these sample data is

$$\eta: \quad \mathbf{y}_s = \mathbf{Z}_s\boldsymbol{\gamma} + \mathbf{X}_s\boldsymbol{\lambda} + \mathbf{e}_{\eta s}$$

where $\mathbf{e}_{\eta s} \sim (\mathbf{0}, \sigma_\eta^2 \mathbf{I}_n)$, $\sigma_\eta^2 \le \sigma_\xi^2$, and we do have $\mathbf{e}_{\eta s} \perp \left(\mathbf{R}_S, \mathbf{I}_U, \mathbf{Z}_U\right)$. Since

$$E_\eta\left(\mathbf{y}_U \middle| \mathbf{Z}_U\right) = E_\xi\left(\mathbf{y}_U \middle| \mathbf{Z}_U\right) = \mathbf{Z}_U\boldsymbol{\beta}$$

it follows that

$$E_\eta\left(\mathbf{y}_s\left|\mathbf{Z}_s\right.\right)=\mathbf{Z}_s\boldsymbol{\gamma}+E_\eta\left(\mathbf{X}_s\left|\mathbf{Z}_s\right.\right)\boldsymbol{\lambda}=\mathbf{Z}_s\boldsymbol{\beta}$$

so η can be re-expressed in the form

$$\eta:\quad \mathbf{y}_s=\mathbf{Z}_s\boldsymbol{\gamma}+E_\eta\left(\mathbf{X}_s\left|\mathbf{Z}_s\right.\right)\boldsymbol{\lambda}+\left\{\mathbf{X}_s-E_\eta\left(\mathbf{X}_s\left|\mathbf{Z}_s\right.\right)\right\}\boldsymbol{\lambda}+\mathbf{e}_{\eta s}=\mathbf{Z}_s\boldsymbol{\beta}+\tilde{\mathbf{X}}_s\boldsymbol{\lambda}+\mathbf{e}_{\eta s}.$$

We can approximate $\tilde{\mathbf{X}}_s$ via Gram-Schmidt orthogonalization, i.e. we put

$$\tilde{\mathbf{X}}_s=\left[\mathbf{I}_s-\mathbf{Z}_s\left(\mathbf{Z}_s^T\mathbf{Z}_s\right)^{-1}\mathbf{Z}_s^T\right]\mathbf{X}_s$$

from which it follows that

$$\tilde{\mathbf{X}}_s^T\mathbf{w}_{\xi s}=\tilde{\mathbf{X}}_s^T\mathbf{1}_n+\underbrace{\tilde{\mathbf{X}}_s^T\mathbf{Z}_s}_{=0}\left(\mathbf{Z}_s^T\mathbf{Z}_s\right)^{-1}\left(\mathbf{Z}_U^T\mathbf{1}_N-\mathbf{Z}_s^T\mathbf{1}_n\right)=\tilde{\mathbf{X}}_s^T\mathbf{1}_n.$$

Optimal weights for the respondent sample under η are therefore

$$\mathbf{w}_{\eta o}=\mathbf{w}_{\xi o}+\tilde{\mathbf{X}}_o\left(\tilde{\mathbf{X}}_o^T\tilde{\mathbf{X}}_o\right)^{-1}\left(\tilde{\mathbf{X}}_U^T\mathbf{1}_N-\tilde{\mathbf{X}}_o^T\mathbf{1}_m\right).$$

Replacing $\tilde{\mathbf{X}}_U^T\mathbf{1}_N$ by its sample estimate $\tilde{\mathbf{X}}_s^T\mathbf{w}_{\xi s}=\tilde{\mathbf{X}}_s^T\mathbf{1}_n$ leads to

$$\mathbf{w}_{\eta o}=\mathbf{w}_{\xi o}+\tilde{\mathbf{X}}_o\left(\tilde{\mathbf{X}}_o^T\tilde{\mathbf{X}}_o\right)^{-1}\tilde{\mathbf{X}}_{s-o}^T\mathbf{1}_{n-m}=\mathbf{w}_{\xi o}+\mathbf{u}_{\eta o}.$$

We say that the respondent sub-sample o is balanced if $\mathbf{u}_{\eta o}=\mathbf{0}$. Adaptive design then corresponds to selecting a subsample of non-respondents for follow-up if o is not sufficiently balanced (or there are not enough respondents).

The process of follow-up of non-respondents can be operationalized as follows. Let f denote a potential follow-up subsample of size k. The aim is to choose f in order to minimise

$$\theta_f=E_\eta\left(\mathbf{w}_{\xi(o+f)}^T\mathbf{y}_{o+f}-\mathbf{w}_{\eta(o+f)}^T\mathbf{y}_{o+f}\right)^2\propto\mathbf{1}_{n-m-k}^T\tilde{\mathbf{X}}_{s-o-f}\left(\tilde{\mathbf{X}}_o^T\tilde{\mathbf{X}}_o\right)^{-1}\tilde{\mathbf{X}}_{s-o-f}^T\mathbf{1}_{n-m-k}.$$

Put $f=\{i\}$. We then have an index $\theta_{\{i\}}$ that allows non-respondents to be ordered so that the most influential can be targeted. The extension of this idea to clustered populations and multiple call-back waves is straightforward.

References

Basu, D. (1971), "An Essay On The Logical Foundations Of Survey Sampling I", in *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.

Brewer, K.R.W. (1963), "Ratio Estimation And Finite Populations: Some Results Deducible From The Assumption Of An Underlying Stochastic Process", *Australian Journal of Statistics*, 5, pp. 93-105.

Chambers, R.L., Dorfman, A.H. & Wehrly, T.E. (1993), "Bias Robust Estimation In Finite Populations Using Nonparametric Calibration", *Journal of the American Statistical Association*, 88, pp. 268-277.

Clark, R.G. & Chambers, R.L. (2008), "Adaptive Calibration For Prediction Of Finite Population Totals", *Survey Methodology*, 34, pp. 163-172.

Deville, J.-C. (1992), "Constrained Samples, Conditional Inference, Weighting: Three Aspects Of The Utilisation Of Auxiliary Information", *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, Statistics Sweden, Örebro, October 5-7 1992.

Deville, J.C. & Särndal, C.E. (1992), "Calibration Estimators In Survey Sampling", *Journal of the American Statistical Association*, 87, pp. 376-382.

Deville, J.-C. & Tillé, Y. (2004), "Efficient Balanced Sampling: The Cube Method", *Biometrika*, 91, pp. 893–912.

Hansen, M.H., Madow, W.G. & Tepping, B.J. (1983), "An Evaluation Of Model-Dependent And Probability-Sampling Inferences In Sample Surveys", *Journal of the American Statistical Association*, 78, pp. 776-793.

Gelman, A. (2007), "Struggles With Survey Weighting And Regression Modeling", *Statistical Science*, 22, pp. 153-164.

Godambe, V.P. (1955), "A Unified Theory Of Sampling From Finite Populations", *Journal of the Royal Statistical Society Series B*, 17, pp. 269-278.

Karmel, T.S. & Jain, M. (1987), "Comparison Of Purposive And Random Sampling Schemes For Estimating Capital Expenditure", *Journal of the American Statistical Association*, 82, pp. 52-57.

Kendall, M. (1959), "Hiawatha Designs An Experiment", *The American Statistician*, 13, pp. 23-24.

Little, R.J.A. (2006), "Calibrated Bayes: A Bayes/Frequentist Roadmap", *American Statistician*, 60, pp. 213-223.
Little, R.J.A. (2012), "Calibrated Bayes: An Alternative Inferential Paradigm For Official Statistics", *Journal of Official Statistics*, 28, pp. 309-372.
Neyman, J. (1934), "On The Two Different Aspects Of The Representative Method: The Method Of Stratified Sampling And The Method Of Purposive Selection", *Journal of the Royal Statistical Society*, 97, pp. 558-606.
Pfeffermann, D. (1993), "The Role Of Sampling Weights When Modelling Survey Data", *International Statistical Review*, 61, pp. 317-337.
Royall, R.M. (1976a), "Current Advances In Sampling Theory: Implications For Human Observational Studies", *American Journal of Epidemiology*, 104, pp. 463-474.
Royall, R.M. (1976b), "The Linear Least Squares Prediction Approach To Two-Stage Sampling", *Journal of the American Statistical Association*, 71, pp. 657-664.
Smith, T.M.F. (1983), "On The Validity Of Inferences From Non-Random Samples", *Journal of the Royal Statistical Society Series A*, 146, pp. 394-403.
Sugden, R.A. & Smith, T.M.F. (1984), "Ignorable And Informative Designs In Survey Sampling Inference", *Biometrika*, 71, pp. 495-506.
Tam, S.M. & Chan, N.N. (1984), "Screening Of Probability Samples", *International Statistical Review*, 52, pp. 301-308.
Welsh, A.H. & Wiens, D.P. (2013), "Robust Model-Based Sampling Designs", *Statistics and Computing*, 23, pp. 689-701.