Pre-Sampling Model Based Inference V

Stephen Woodruff

Specified Designs, 800 West View Terrace, Alexandria, VA 22301

Keywords: Design Based Inference, Model Based Inference, Pre-sampling Model Based Inference

Abstract

Randomized construction of population units is combined with randomized sampling of these population units to provide a more complete description of the stochastic structure of sample data. This additional structure imposes a model on population units` data, a model that is a consequence of this additional structure and thus with the same credibility as probabilities of selection in Design Based Inference. This expanded theory is called, Pre-Sampling Model Based (PSMB) Inference. It eliminates problems with both design effect and model fit. The result can be estimators with much smaller sampling error than Design Based Estimators. PSMB estimators and design based estimators are evaluated with respect to sampling error from repeated sampling of population units under stratified cluster designs. Causes of increased error in Design Based Inference are noted.

1. Introduction and Conclusions

This paper suggests that for many populations it could be very informative to expand the basis for sample survey inference from randomized sample selection to both randomized sample selection and randomized unit construction. The definition of a population unit often implies that the unit came about through implicit randomization as in the case of sampling a body of water to measure prevalence of quantifiable characteristics (like variables related to pollution). In this case the unit would be the container of water and its contents may often be reasonably described as a random sample from the entire body.

Randomized sample unit construction is an important analytic tool in survey inference for such populations. Population units can viewed as random samples of smaller entities called atoms and the unit study variables are the sum over the unit`s atoms of the same study variables attached to each atom in the unit. In vector notation let V_{ij} be the vector of study variables attached to the j^{th} atom of the i^{th} unit (sample or universe), then if V_i is the vector of study variables attached to the i^{th} unit,

$V_i = \sum_{j \epsilon S_i} V_{ij}$, where S_i is the set of atoms in the i^{th} unit.

The population of atoms is partitioned randomly into units. Every atom in the population is in one and only one unit and thus the population sum over the atoms of the atom study variables is the same as the sum over the population of units of the unit study variables.

Although randomized unit construction is used to impose a model on the population, this element of randomization is not directly a part of sampling error. Sampling error is defined as in Design Based inference, with respect to sample unit selection, usually under stratified cluster designs where moments of sample estimates are defined conditional on the outcomes of randomized unit construction and with respect to repeated outcomes of

sample unit selection. Randomized unit construction leads to an efficient estimator that controls both bias and variance where Design Based estimators although unbiased, can have large variances under inefficient sample designs often encountered in practice.

Randomized unit construction also provides more informative variance expressions under repeated unit sampling by introducing additional population parameters to sampling theory. Randomized unit construction imposes a model on population units, a model not dependent on observed data but rather dependent on the random process that produces the unit`s data. This method of modeling eliminates a criticism of model based inference when a model is conjectured from available data on the study population, data that may be either stale or from a sample that could be anomalous.

The process of randomized unit construction is called Pre-sampling. Under the model derived from Pre-sampling, there is a Best Linear Unbiased Estimator (BLUE). This estimator is particularly useful in cases where design effect seriously increases the variance of standard Design Based estimators. Although Design Based inference provides unbiased estimates the particular sample design may substantially increase the variance of these estimates.

PSMB models come with the same level of credibility as probabilities of selection in design based inference, but PSMB inference also controls variance while Design Based inference does not, unit sample designs tend to be administratively defined rather than statistically optimized.

The goals of PSMB inference are:

1) Inference from sample data should be based randomization. Randomization gives access to probability theory as the tool of analysis and provides a certain comforting impartiality.
2) It should minimize de-randomization, a term which refers to post sampling adjustments that are ubiquitous in Design Based inference. De-randomization includes non-response adjustment, outlier adjustments, and other procedures that bring survey estimates in line with anticipated results. Better explore inference methodologies that need minimal post sampling adjustment.
3) It should recognize and exploit the multivariate structure of most sample and population data, its covariance structure, and the large quantity of relevant data usually available. When Design Based sampling inference was developed the computing power was not available to store or analyze the quantity of relevant data currently available.
4) Model Based procedures founded on randomization are derived from Pre-sampling. The models derived this way provide the hypothesis for Best Linear Unbiased Estimation (Estimators that control both bias and variance under repeated sampling from complex stratified cluster designs particularly when Design Based estimators suffer severe design effects).

The following table summarizes the main advantages and disadvantages of Design Based inference, Model Based inference, and PSMB inference. PSMB inference keeps the advantages of both Design Based inference and Model Based inference while avoiding their respective disadvantages, see Table 1.

Table 1. Important Properties of Inference Methodologies

Inference Methodology	Advantages	Disadvantages
Design Based Inference	**-Unbiased Estimates -Probabilities of selection imposed by randomization**	**-Only Unbiased Estimates -Variance is dependent on sample design**
Model Based Inference	**-Both Unbiased & Variance Controlled Estimates**	**Model Conjectured from either stale historical population data or current sample data.**
PSMB Inference	**-Model imposed by randomization -Both Unbiased & Variance Controlled Estimates**	**Ø**

Table 2. Some examples of populations where population units are well approximated as random samples of atoms

UNITS	ATOMS	STUDY VARIABLES
Containers of mail pieces	Individual mail pieces	Weight, postage, counts
Boxes of fruit	Individual pieces of fruit	Weight, counts by type (damaged or whole)
Business Establishments	Individual employees	Hours worked, wages, benefits (etc)
Blood samples	Individual blood cells	Counts by type,
Fields of corn	Individual ears of corn	Yield, weight, calories, etc.
Containers of whatever	Individual items within container	Weight, counts by type, any quantitative variable(s).

Randomized construction of population units (Pre-sampling) can model unit heterogeneity through atom homogeneity within atom type, by varying the numbers of different atom types within a unit. This atom structure also provides a good description of the covariance structure of unit study variables that is particular to each unit in the population, depending upon a unit`s mix of atom types.

Design Based variance expressions under repeated sampling are more informative when they are expressed in terms of atom Pre-sampling model parameters, Woodruff (2011). These expressions help to clarify situations where Design Based inference will be problematic prior to actually seeing these Design Based estimates. Many survey organizations seem to have a survey or two where Design Based Inference regularly produces estimates which appear strange to subject matter experts.

The theme of this paper is that sampling theory can be enhanced by expanding it from randomized selection of sample units to both their randomized selection and their randomized construction (or synthesis). The details of PSMB inference are presented in JSM proceedings papers by Woodruff from 2006 to 2012. For many sample designs encountered in practice the PSMB estimates will have orders of magnitude smaller

repeated unit sampling variance than standard design based estimates. And in general, the PSMB estimates will have smaller variance although only slightly smaller when the unit sampling design is efficient for Design Based inference. In the case of stratified cluster designs this generally will be the case when clusters are similar in size and strata ratios are also similar to one another.

Woodruff (2011) contains the derivation of expectation and variance expressions of PSMB and Design Based estimators under repeated sampling from stratified cluster designs. Woodruff (2012) applies PSMB inference to common sampling problems; variance estimates are derived and comparisons of Design Based and PSMB estimators derived. Woodruff (2009) proposes variance estimation methodologies for PSMB estimators and describes several simulation studies to evaluate them. Other papers Woodruff (2010, 2009, 2008, 2007, and 2006) derive PSMB estimators under varying degrees of generality and describe simulation studies where they are tested. These papers outline techniques that can be applied to PSMB applications. It still remains to derive informative variance expressions for PSMB estimators under more generality - more than one auxiliary variable.

In conclusion, randomized construction of population units is combined with randomized sampling of these population units to provide a more complete description of the stochastic structure of sample survey estimates. This additional structure imposes a model on population units` data, a model that is a consequence of randomization, and a model with the same credibility as probabilities of selection in Design Based Inference. PSMB inference is one way to achieve the four goals listed above for sampling inference.

References

Woodruff, S. M. (2006), "Probability Sample Designs that Impose Models on Survey Data", Proceedings of the American Statistical Association, Survey Research Methods

Woodruff, S. M. (2007), "Properties of the Combined Ratio Estimator and a Best Linear Unbiased Estimator When Design Control is Problematic", Proceedings of the American Statistical Association, Survey Research Methods

Woodruff, S. M. (2008), "Inference in Sampling Problems Using Regression Models Imposed by Randomization in the Sample Design - Called Pre-Sampling", Proceedings of the American Statistical Association, Survey Research Methods

Woodruff, S. M. (2009), "An Introduction to Pre-Sampling Inference I" Proceedings of the American Statistical Association, Survey Research Methods

Woodruff, S. M. (2010), "An Introduction to Pre-Sampling Inference II" Proceedings of the American Statistical Association, Survey Research Methods

Woodruff, S. M. (2011), "An Introduction to Pre-Sampling Inference III" Proceedings of the American Statistical Association, Survey Research Methods

Woodruff, S. M. (2012), "An Introduction to Pre-Sampling Inference IV" Proceedings of the American Statistical Association, Survey Research Methods