Statistical Theory and Related Fields

On valid descriptive inference from non-probability sample

--Manuscript Draft--

Full Title:	On valid descriptive inference from non-probability sample
Manuscript Number:	TSTF-2018-0039R3
Article Type:	Original Article
Keywords:	non-informative selection; prediction model; calibration; inverse propensity weighting; sample matching; model misspecification
Abstract:	We examine the conditions under which descriptive inference can be based directly on the observed distribution in a non-probability sample, under both the super-population and quasi-randomisation modelling approaches. Review of existing estimation methods reveals that the traditional formulation of these conditions may be inadequate due to potential issues of under-coverage or heterogeneous mean beyond the assumed model. We formulate unifying conditions that are applicable to both type of modelling approaches. The difficulties of empirically validating the required conditions are discussed, as well as valid inference approaches using supplementary probability sampling. The key message is that probability sampling may still be necessary in some situations, in order to ensure the validity of descriptive inference, but it can be much less resource-demanding provided the presence of a big non-probability sample.
Order of Authors:	Li-Chun Zhang
Response to Reviewers:	last paragraph of section 3.2 on p. 13 revised

On valid descriptive inference from non-probability sample

Li-Chun Zhang[1]

Abstract: We examine the conditions under which descriptive inference can be based directly on the observed distribution in a non-probability sample, under both the super-population and quasi-randomisation modelling approaches. Review of existing estimation methods reveals that the traditional formulation of these conditions may be inadequate due to potential issues of under-coverage or heterogeneous mean beyond the assumed model. We formulate unifying conditions that are applicable to both type of modelling approaches. The difficulties of empirically validating the required conditions are discussed, as well as valid inference approaches using supplementary probability sampling. The key message is that probability sampling may still be necessary in some situations, in order to ensure the validity of descriptive inference, but it can be much less resource-demanding given the presence of a big non-probability sample.

Keywords: non-informative selection, prediction model, calibration, inverse propensity weighting, sample matching, model misspecification

1 Introduction

There is a resurgence of interest in the use of non-probability samples. See e.g. Baker et al. (2013) and Elliot and Valliant (2017) for two recent reviews. Such data may arise in situations where probability sampling is either infeasible or too costly. The observations may be obtained from the so-called big-data sources, such as payment transaction data via a specific platform, cellphone call data from a major provider of the service. These big-data non-probability samples can be much larger in size, compared to the more familiar non-probability samples collected from web panel surveys, quota sampling, etc.

Following Rubin (1976) and Little (1982), Smith (1983) considers the so-called *super-population (SP)* approach to inference from non-probability sample. Under this approach, a prediction model is constructed for the outcome variable of interest, often conditional on some chosen covariates. In particular, Smith (1983) observes an important distinction between *analytic* and *descriptive* inference. In analytic inference, the target is the model parameters that are of a theoretical nature; such parameters can never be observed directly no matter how large the sample is. Whereas the targets of descriptive inference are statistics of a given finite population, such that in principle they can be directly observed given a perfect census of the population.

Moreover, Smith (1983) focuses on *validity conditions*, under which the non-probability sample observation mechanism can be ignored, in the sense that inference can be based

[1] Address for correspondence: S3RI/University of Southampton, Highfield SO17 1BJ, Southampton, UK. Email: L.Zhang@soton.ac.uk

on the *observed* distributions directly, such as the conditional distribution of the outcome variable given the covariates in the sample. The two key validity conditions under the SP approach can be roughly stated as follows: (i) the prediction model is correctly specified for the population units, (ii) the non-probability sample selection mechanism is non-informative, in the sense that the relevant distribution under the population model can be observed in the non-probability sample directly. Similar validity conditions for the SP approach apply in other situations, such as purposive sampling (Royall, 1970), missing data problems (Rubin, 1976).

In this paper we concentrate on descriptive inference methods that depend on validity conditions in the sense of Smith (1983). Of course, inference is also possible without such validity conditions. For instance, not missing-at-random models (Rubin, 1976) can be used to deal with informative missing data, where the unobserved full-sample outcome distribution is not the same as that among the respondent subsample. Or, the sample likelihood of Pfeffermann et al. (1998) can be applied to survey data under informative sampling, where the distribution that holds in the population cannot be directly observed in the sample. See also Pfeffermann (2017) for several other situations where this approach may be relevant. We do not consider such approaches here, which require explicitly modelling the informative observation mechanism of sample selection or measurement.

As reviewed by Elliot and Valliant (2017), there exists another *quasi-randomisation (QR)* approach to non-probability samples. Under the QR approach, one hypothesises a randomisation model of the non-probability sample inclusion indicator, but treats the outcomes of interest as unknown constants in the population. Though it is clearly inspired by the randomisation approach based on probability sampling, the QR approach is also a model-based approach, based on a model of the sample inclusion indicator instead of a prediction model of the outcome variable under the SP approach. A key motivation is that the correct inclusion probability can be used for any outcome of interest, just like when it is known under probability sampling, whereas the SP approach by nature must be specified differently for different outcome variables. In the context of survey sampling, the QR approach was introduced to deal with nonresponse, where response to survey is modelled as the second phase of selection, in addition to the first phase of sample selection according to a probability sampling design (Oh and Scheuren, 1983).

According to Elliot and Valliant (2017), two key validity conditions are required for the QR approach. (I) The non-probability sample does have a probability sampling mechanism, even though it is unknown. In particular, one assumes that this hypothesised sample inclusion probability is strictly positive for all the population units, so that the only difference to probability sampling is that the inclusion probability is unknown. (II) There exist a set of covariates that "fully govern the sampling mechanism". In other words, the sample inclusion probability is a function of these covariates.

Thus, there are two model-based approaches to inference from non-probability sample. Under the SP approach, one models the outcome variable conditional on the realised sample inclusion indicators; whereas under the QR approach, one models the sample inclusion indicators, but treats the outcomes as unknown constants. Although one may envisage the outcomes as the realised values of random variables, a fully specified model of the outcome variable will not be required under the QR approach, given suitable validity conditions. Similarly, although one acknowledges that the sample selection mechanism may be critical to the SP approach, a fully specified model of the inclusion indicator will not be required under the SP approach, given suitable validity conditions.

It is possible to construct estimators that combine both the models of outcome and sample inclusion indicator, in a manner such that the estimator is consistent as long as one of the two models hold. Over the recent years, it is becoming common to refer to this estimation approach as "doubly robust" (Robins et al., 1994; Kang and Shafer, 2007; Kim and Haziza, 2014). Notice that the traditional generalised regression estimator in survey sampling is doubly-robust in the same sense, except that here the randomisation mechanism is actually known. Nevertheless, it is a fact that in the debate between model-based and design-based inference from probability sampling, either side questioned the "robustness" of the other.

The rest of the paper is organised as follows. In Section 2 we review the estimation methods from non-probability sample which do require validity conditions. Although these have been roughly stated above, a closer examination under both modelling perspectives reveals nuances across the different estimators. Moreover, we shall highlight the potential challenges of under-coverage and heterogeneous means beyond the assumed model. The traditional formulation of validity conditions is inadequate in both regards. We outline a set of unified validity conditions in Section 3, which are formulated non-parametrically and encompasses both the modelling approaches. Post-stratification and calibration estimators are considered in light of these conditions. However, as will be discussed, a key difficulty in practice is that the validity conditions may be impossible to verify empirically based only on the data used for the estimation. Finally, we outline shortly in Section 4 some valid approaches given a supplementary probability sampling of the outcome of interest, followed by a brief summary in Section 5.

The key message is that probability sampling may still be necessary in some situations, in order to ensure the validity of descriptive inference, but it can be much less resource-demanding given the presence of a big non-probability sample. In fact, the bigger the non-probability sample, the better it is.

2 Review of existing approaches

Denote by U the population of known size N. Let each population unit be associated with an *outcome* of interest, denoted by y_i, for $i \in U$. Denote by B the *observed* nonprobability sample of size n_B. A common assumption to all the estimators we discuss below is that B does not contain any out-of-scope units, such that $B \subset U$, and there are no duplicated units in B. Let $\delta_i = 1$ if $i \in B$, and 0 if $i \in U \setminus B$. Let y_i be observed for all the units in B, and let $y_B = \{y_i; i \in B\}$. To fix the idea, let

$$Y = \sum_{i \in U} y_i$$

be the population total that is the target of descriptive inference. Let $x_B = \{x_i; i \in B\}$ in cases where any relevant covariates x_i are available in the sample B. Let $X = \sum_{i \in U} x_i$ be the population totals and let $\bar{X} = X/N$. Given x_B, one can have two situations depending on whether $(X, \bar{X})$ are known or not. In the case they are unknown, it may still be possible that there exists a second *probability sample* S, for $S \subset U$, in which x_i is observed, so that $(X, \bar{X})$ can be estimated based on the sample S.

2.1 B-sample expansion estimator

Consider first the most basic situation where only y_B is observed, and no relevant covariates are available at all. Let $\bar{y}_B = \sum_{i \in B} y_i / n_B$ be the B-sample mean. The B-sample expansion estimator of Y is given by

$$\widehat{Y} = N \bar{y}_B \tag{1}$$

Under the SP approach, let

$$\mu_i = E(y_i | \delta_i, i \in U)$$

be the conditional expectation of y_i given δ_i, for any $i \in U$, where both δ_i and y_i are treated as random variables. Provided the conditional expectation is the same as the unconditional expectation given either $\delta_i = 1$ or $\delta_i = 0$, for any $i \in U$, denoted by

$$\mu = \mu(\delta_i = 1) = E(y_i | i \in B) = E(y_i; i \in U) \tag{2}$$

we have

$$E(\bar{y}_B - Y/N | B) = \sum_{i \in B} \mu / n_B - \mu = 0$$

such that the B-sample expansion estimator is prediction unbiased for Y. We shall refer to (2) as the *SP assumption*, which is a validity condition for the B-sample expansion estimator under the SP approach.

Under the QR approach, where y_i is treated as a fixed constant, let

$$p_i = \Pr(\delta_i = 1; y_i, i \in U)$$

be the *inclusion probability* of any population unit that is associated with the value y_i. The notation ";" is used here instead of "|" because, strictly speaking, p_i is not a conditional probability now that y_i is not conceived as the realised value of a random variable under the QR approach. Now, provided the inclusion probability is the same for any $i \in U$,

$$p_i = p \tag{3}$$

we have $\widetilde{Y} = \sum_{i \in B} y_i/p$ is unbiased for Y, since

$$E(\sum_{i \in B} y_i/p) = \sum_{i \in U} E(\delta_i; y_i, i \in U) y_i/p = \sum_{i \in U} p y_i/p = Y$$

In reality, p is unknown. It is natural to estimate it by $\hat{p} = n_B/N$ under (3), which yields (1) as the resulting plug-in estimator. It follows that the *QR assumption* (3) is the key validity condition, which ensures that the B-sample expansion estimator is consistent for Y, as $N \to \infty$ and $n_B/N = O_p(1)$ asympotically.

In summary, the B-sample expansion estimator (1) can be motivated under both the SP and QR approaches, given validity conditions (2) and (3), respectively.

2.2 B-sample calibration estimator

Suppose relevant covariates x_B are available in the sample B. The population totals X may be either known or unknown. In the latter case, suppose they can be estimated from a second probability sample S. The B-sample calibration estimator of Y is given by

$$\widehat{Y} = \sum_{i \in B} w_i y_i \qquad \text{where} \quad \begin{cases} \sum_{i \in B} w_i x_i = X & \text{if known } X \\ \sum_{i \in B} w_i x_i = \widehat{X}(S) & \text{if unknown } X \end{cases} \tag{4}$$

where $\widehat{X}(S)$ is some consistent S-sample estimator, as the S-sample size increases, and the weights $w_B = \{w_i; i \in B\}$ are calibrated in a way depending on the availability of X.

To actually compute the estimator (4), one needs to choose a set of initial weights, denoted by $a_B = \{a_i; i \in B\}$, and a distance function such as $\sum_{i \in B}(w_i - a_i)^2/a_i$ between the initial and calibrated weights (Deville and Särndal, 1992). In the case of

$$a_i = 1/p_i \tag{5}$$

where p_i is the true B-sample inclusion probability, for $p_i > 0$, the calibration estimator is consistent, as $N \to \infty$ and $n_B/N = O_p(1)$, given mild regularity conditions in addition. However, insofar as one cannot manage to set the initial weights (5), the calibration estimator is unmotivated from the QR perspective.

Next, under the SP approach, suppose the SP$_x$ assumption given by

$$E(y_i | x_i, i \in U) = \mu(x_i) = x_i^\top \beta \tag{6}$$

which relates the conditional expectation of y_i linearly to the given x_i, and

$$E(y_i|x_i, i \in U) = E(y_i|x_i, i \in B) \tag{7}$$

by which the B-sample selection is *non-informative* given x_i. We have then

$$E(\sum_{i \in B} w_i y_i - Y|x_U) = E(\sum_{i \in B} w_i x_i^\top \beta) - X^\top \beta = 0$$

given $\sum_{i \in B} w_i x_i = X$, regardless of the initial weights a_B. Otherwise, this expectation would tend to 0, provided $\widehat{X}(S)$ is an asymptotically unbiased estimator of X, under some suitable asymptotic setting. It follows that the assumptions (6) and (7) are the key validity conditions for the B-sample calibration estimator under the SP approach.

The estimator (4) becomes the B-sample post-stratification estimator in the special case where x_i is the post-stratum dummy index. For the QR approach, one can set a_i to be the inverse post-stratum B-sample fraction, which is equivalent to introducing the QR assumption (3) in each post-stratum separately. This *QR_x assumption* provides then a validity condition for the B-sample post-stratification estimator under the QR approach. For the SP approach, the two assumptions (6) and (7) remain formally the same.

2.3 B-sample inverse propensity weighting

Suppose relevant covariates x_B are available in the sample B. The B-sample inverse propensity weighting (IPW) estimator is constructed under the QR approach. Suppose

$$p_i = p(x_i; \eta) > 0 \tag{8}$$

i.e. the B-sample inclusion probability is completely determined given x_i, in the strictly positive parametric form $p(x_i; \eta)$, which may as well be referred to as the *QR_x assumption*. Provided x_i is known for all the units in the population, η can be estimated, say, by a population estimating equation

$$\sum_{i \in U} H(\delta_i; \eta) = 0$$

where $E[H(\delta_i; \eta)] = 0$. Otherwise, suppose x_S is observed in a second probability sample S, one can use the pseudo population estimating equation

$$\sum_{i \in S} d_i H(\delta_i; \eta) = 0$$

(Kim and Wang, 2018), where d_i is the sampling weight, for $i \in S$, or some S-sampling design-consistent adjustment of it. This requires that one is able to observe δ_i for each unit i in S, in other words the two samples S and B can be matched, which is an important assumption in terms of application. To ensure that $H(\delta_i; \eta)$ is the same in both of these two estimating equations, i.e. whether $i \in S$ or just $i \in U$, one needs to assume that S-sampling from U is non-informative for δ_i, so that

$$\Pr(\delta_i = 1 | x_i, i \in S) = \Pr(\delta_i = 1 | x_i, i \in U) \tag{9}$$

Notice that, given non-informativeness (9), we have $E[H(\delta_i; \eta)] = 0$ for all $i \in s$, such that one can also use the unweighted S-sample estimating equation, which is given by

$$\sum_{i \in S} H(\delta_i; \eta) = 0$$

instead of the pseudo population estimating equation. Having obtained the parameter estimate $\hat{\eta}$, one obtains $\hat{p}_i = p(x_i; \hat{\eta})$ and the B-sample IPW estimator

$$\widehat{Y} = \sum_{i \in B} y_i / \hat{p}_i \tag{10}$$

which is consistent for Y under mild regularity conditions, if $\hat{\eta}$ is consistent for η under some suitable asymptotic setting. It follows that the QR$_x$ assumption (8) is its key validity condition, whereas the non-informativeness assumption (9) is needed in addition when x_i is only available in a probability sample S instead of the population.

2.4 Another B-sample IPW estimator

Elliot and Valliant (2017) discuss another IPW estimator (10), where p_i is obtained with the help of a second so-called reference probability sample S, and is given by

$$p_i \propto \Pr(S_i = 1 | x_i, i \in U) \frac{\Pr(\delta_i = 1 | x_i, i \in B \cup S)}{\Pr(S_i = 1 | x_i, i \in B \cup S)} \tag{11}$$

where $S_i = 1$ if $i \in S$ and 0 if $i \in U \setminus S$, and to fix the idea one may suppose $S \cap B = \emptyset$. Firstly, the QR$_x$ assumption (8) is retained. The definition of p_i by (11) can then be motivated as follows:

$$\begin{aligned}
\frac{\Pr(\delta_i = 1 | x_i, i \in U)}{\Pr(S_i = 1 | x_i, i \in U)} &\propto \frac{\Pr(x_i | \delta_i = 1, i \in U)}{\Pr(x_i | S_i = 1, i \in U)} \qquad \left[\text{prop. to } \frac{\Pr(\delta_i = 1 | i \in U)}{\Pr(S_i = 1 | i \in U)}\right] \\
&\propto \frac{\Pr(x_i | \delta_i = 1, i \in B \cup S)}{\Pr(x_i | S_i = 1, i \in B \cup S)} \\
&\propto \frac{\Pr(\delta_i = 1 | x_i, i \in B \cup S)}{\Pr(S_i = 1 | x_i, i \in B \cup S)} \qquad \left[\text{prop. to } \frac{\Pr(\delta_i = 1 | i \in B \cup S)}{\Pr(S_i = 1 | i \in B \cup S)}\right]
\end{aligned}$$

provided the S-sample inclusion probability is also fully determined by x_i in the sense of (8). Thus, the validity condition for the IPW estimator (10) based on (11) is that the QR$_x$ assumption (8) holds for *both the samples, given the same* x_i.

We make two observations. Firstly, despite the superficial resemblance to the propensity scoring method of Rosenbaum and Rubin (1983), the above argument for p_i is not the same. As Rosenbaum and Rubin (1983) state clearly before their first enumerated equation, "In this paper, the N units in the study are viewed as a simple random sample from some population", where N is the size of the combined sample of treatment and non-treatment. The analogy to this combined sample is $B \cup S$ here. However, it is generally untenable that $B \cup S$ can be treated as a simple random sample from the population. Secondly, for any given probability sample S, it is possible to identify the variables that determine the designed inclusion probability, denoted by $\pi_i = \pi(z_i)$, for $i \in U$. There arises thus a question, "what if $\pi(z_i)$ differs considerably from $p(x_i, \hat{\eta})$?" Moreover, one may have more than one probability sample in which x_i is observed. There arises then a question, "which reference sample should one use?"

2.5 Sample matching estimator

Rivers (2007) applies the SP approach in situations where a second probability sample S is available. Yang and Kim (2018) study mass imputation methods, which include the matching estimator of Rivers (2007) as a special case. The sample matching (SM) estimator is given by

$$\widehat{Y} = \sum_{i \in S} d_i \hat{y}_i \tag{12}$$

where $\hat{y}_i = y_{k_i}$, for $k_i = \arg\min_{j \in B} \|x_i - x_j\|$ based on a chosen metric $\|\cdot\|$. That is, y_{k_i} is the nearest-neighbour (NN) imputation value from the B-sample for $i \in S$.

To focus on the idea, assume for the moment exact matching is the case, where $x_{k_i} = x_i = x$ for all $i \in S$ and $k_i \in B$. We have then $E(\hat{y}_i | x_i = x) = E(y_{k_i} | x_{k_i} = x, k_i \in B)$, which is the same as $E(y_i | x_i = x, i \in B)$ as if the unit i were in B. Given the non-informativeness assumption (7) for the B-sample, which Yang and Kim (2018) call the "ignorability" assumption, we have

$$E\big[\sum_{i \in S} d_i E(\hat{y}_i | x_i)\big] = E\big[\sum_{i \in S} d_i E(y_i | x_i, i \in B)\big] = E\big[\sum_{i \in S} d_i E(y_i | x_i, i \in U)\big]$$
$$= \sum_{i \in U} E(y_i | x_i, i \in U) = E(Y | x_U)$$

With respect to both the population model and the design of S, the SM estimator (12) is prediction unbiased for Y. Notice that in the case of $S = U$, the SM estimator is just an NN-imputation method. Whether $S = U$ or not, the NN-imputed SM estimator is likely to be less efficient than a prediction-imputed SM estimator

$$\hat{Y} = \sum_{i \in S} d_i E\big(y_i | x_i; \widehat{\beta}(B)\big)$$

whenever a correct parametric specification of the conditional mean (via β) is possible.

The simulations results of Yang and Kim (2018) show that NN-imputation is less efficient than imputation based on semi-parametric generalised additive models.

Now, it is not difficult to see that the consistency of the SM estimator (12) can be established, given asymptotic exact matching instead, i.e.

$$\|x_i - x_{k_i}\| \to 0 \text{ in probability,} \tag{13}$$

for any $i \in S$, as $N \to \infty$ and $n_B/N = O_p(1)$. Yang and Kim (2018) make the assumption of "common support" to the same effect. To ensure that $E(y_i|x_i, i \in U)$ does not change abruptly as the value x_i varies, Yang and Kim (2018) assume that $E(y_i|x_i, i \in U)$ is continuous differentiable. Or, one may adopt the SP$_x$ assumption below:

$$\|E(y_i|x_i, i \in U) - E(y_j|x_j, j \in U)\| = O\big(\|x_i - x_j\|\big) \text{ as } N \to \infty \tag{14}$$

(Chen and Shao, 2000, Theorem 1). It follows that the assumptions (7), (13) and (14) are the key validity conditions for the consistency of the SM estimator (12).

We make two observations. Firstly, an attractive feature of the NN-imputation is that the imputed sample S looks more realistic and natural than, say, by the regression prediction imputation. However, unless the S-sampling is non-informative, the NN-imputed S-sample will not resemble the true S-sample that could have been observed, since

$$E(\hat{y}_i|x_i, i \in S) = E(y_i|x_i, i \in U) \neq E(y_i|x_i, i \in S)$$

where the inequality is the case unless S-sampling is non-informative in the sense of (7). Secondly, for any other covariate $z_i \neq x_i$, including when z_i contains the S-sample design variables, we have

$$E(\hat{y}_i|z_i, x_i, i \in S) = E(y_i|x_i, i \in U) \neq E(y_i|z_i, x_i, i \in U)$$

unless y_i and z_i are conditionally independent of each other given x_i. This is a general problem for statistical matching of variables associated with distinct units, i.e. y_i associated with x_i for some $i \in B$ and z_i associated with the same value x_i but for some different unit in S. The following example illustrates both remarks above.

Example: Let y_i be independent of $x_i \sim \text{Unif}(0,1)$, for any $i \in U$. Then, the SP$_x$ assumption (14) holds trivially, as long as the marginal expectaion $E(y_i)$ exists. Next, suppose simple random sample B, so that the non-informative assumption (7) holds, and $E(\hat{y}_i|x_i, i \in S) = E(y_i|i \in U)$ regardless of the exact matching assumption. Suppose stratified simple random S-sampling with two strata of different sampling fractions, so that the S-sample inclusion probability is not a constant. Then, the S-sampling is informative (given x_i) as long as the population stratum means are different, since

$$E(\bar{y}_S|x_S, S) = E(\bar{y}_S|S) \neq E(\bar{Y}) = E(\bar{Y}|x_U)$$

where $\bar{y}_S$ is the true S-sample mean that is unknown, since y_i is not observed in S. It follows that the NN-imputed S-sample $\{\hat{y}_i; i \in S\}$ would look like a sample generated by simple random sampling, rather than the actual stratified sampling. Moreover, the SM-estimator of stratum means, corresponding to say $z_i = 1, 2$, respectively, will be biased for the population stratum means.

3 More generally on validity conditions

Non-informative selection in form of (7) or (9) is a critical condition for all the methods in Section 2, which make use of auxiliary variable x_i. Two kinds of possible violation of these assumptions are worth noting.

First, Kim and Rao (2018) point out an important issue that has not received sufficient attention in these methods, namely B-sample under-coverage is the case if some population units have in fact zero chance of being included in it. Under the SP approach, extrapolation of the conditional distribution of y_i in the B-sample to these population units can only be based on subjective beliefs but not empirical evidence. The QR approach is equally affected, since randomisation inference would have been invalidated even if p_i were known for all the B-sample units, let alone when it is unknown and needs to be estimated. To address the issue, Kim and Rao (2018) consider a two-phase SM estimator. Let the S-sample be partitioned into S_1 and S_0, such that $S_1 = \{i; p_i > 0\}$ and $S_0 = \{i; p_i = 0\}$. First, estimate this unobserved partition via the B-sample support:

$$\hat{S}_1 = \{i; \min_{j \in B} \|x_i - x_j\| < \epsilon\}$$

Each S-sample unit that is unsupported in the B-sample ϵ-neighbourhood is assigned to $\hat{S}_0$. Let us suppose this partition estimator is consistent in the following sense:

$$|\hat{S}_1 \cup S_1| / |\hat{S}_1 \cap S_1| \to 1 \text{ in probability,}$$

as $N \to \infty$ and $\epsilon \to 0$. Next, the two-phase SM estimator is given as

$$\widehat{Y} = \sum_{i \in \hat{S}_1} d_i w_{2i} \hat{y}_i$$

where $\sum_{i \in \hat{S}_1} d_i w_{2i} x_i = \sum_{i \in S} d_i x_i$. In other words, the under-coverage is dealt with by the calibration of the weights w_{2i}. This can be motivated, provided the conditional mean $E(y_i|x_i, p_i = 0)$ can be linearly related to x_i, and the relationship is the same for the units with $p_i > 0$, i.e. the under-coverage is non-informative for the SP linear model.

Second, insofar one requires either an assumption of SP_x (7) or QR_x (9), there is always the possibility of *heterogeneous mean*, beyond what is controlled by the chosen x_i. Let $U_x = \{i; x_i = x, i \in U\}$ be of the size N_x. Under the SP approach, which models the mean μ_i of unit i by $\mu(x_i)$, heterogeneous mean is the case if $\mu_i \neq \mu(x_i)$, despite

$$\mu(x) = \sum_{i \in U_x} \mu_i / N_x \tag{15}$$

and $\mu(x_i)$ is statistically correct in that the μ_i's average to $\mu(x)$ over all the units in U_x. Under the QR approach, heterogeneous mean is the case if $p_i \neq p(x_i)$, despite

$$p(x) = \sum_{i \in U_x} p_i / N_x \tag{16}$$

Let us illustrate the concept of heterogeneous mean with a simple example.

Example: Let $x \equiv 1$, such that $\mu(x_i) = \mu$, for all $i \in U$. Let $U = U_1 \cup U_0$ be a partition. Let U_1 be of size N_1 and with mean $\mu_i = \mu(1)$, for all $i \in U_1$; let U_0 be of size N_0 and with mean $\mu_i = \mu(0)$, for all $i \in U_0$. Suppose $\mu(1) \neq \mu(0)$. Let $\mu = \mu(1)N_1/N + \mu(0)N_0/N$. Then, $\mu_i \neq \mu$ for any $i \in U$, but we still have $\sum_{i \in U} \mu_i / N = \mu$, satisfying (15).

Heterogeneous mean affects the SP and QR approaches differently. Given (15), assuming $\mu_i = \mu(x)$ for $i \in U_x$ is prediction unbiased, despite heterogeneous mean, since

$$\sum_{i \in U_x} [E(y_i|\delta_i) - \mu(x)] = \sum_{i \in U_x} [\mu_i - \mu(x)] = 0$$

Meanwhile, given (16), assuming $p_i = p(x)$ for $i \in U_x$ yields

$$E\big(\sum_{i \in U_x} \frac{\delta_i y_i}{p(x)}\big) - \sum_{i \in U_x} y_i = p(x)^{-1} \sum_{i \in U_x} \big(p_i - p(x)\big) y_i \neq 0$$

in which case the IPW estimator under the QP approach may be biased, despite the model of p_i is statistically correct in the sense of (16).

The discussion above suggests that the formulation of validity conditions in Section 2 is inadequate in the presence of under-coverage and mean heterogeneity. Below we first reformulate the validity conditions, which cover both the SP and QR approaches, despite the presence of under-coverage and mean heterogeneity. We elaborate and illustrate these conditions for the post-stratification and calibration estimators. Finally, we discuss the difficulties of verifying these validity conditions empirically.

3.1 Non-parametric asymptotic (NPA) non-informativeness

We start by noticing that the B-sample mean equals to the population mean, denoted by $\bar{y}_B = \bar{Y}$, provided

$$\begin{cases} Cov_N(\delta_i, y_i) = \frac{1}{N} \sum\limits_{i \in U} \delta_i y_i - \big(\frac{1}{N} \sum\limits_{i \in U} \delta_i\big)\big(\frac{1}{N} \sum\limits_{i \in U} y_i\big) = 0 \\ E_N(\delta_i) = \sum\limits_{i \in U} \delta_i / N > 0 \end{cases}$$

where E_N and Cov_N denote, respectively, expectation and covariance with respect to the empirical distribution function that places point mass $1/N$ on each population unit. This provides an empirical formulation of the non-informativeness of the B-sample observation mechanism with respect to the outcome of interest. Similar expressions have appeared in various discussions of the potential sample mean bias due to the observation mechanism, such as unequal probability sampling (Rao, 1966), survey nonresponse (Bethlehem, 1988), or big data (Meng, 2018). It motivates the following *non-parametric asymptotic (NPA)* non-informativeness assumption in the absence of any covariates:

$$\begin{cases} \lim\limits_{N \to \infty} Cov_N(\delta_i, y_i) = 0 & \text{i.e. non-informative B-selection} \\ \lim\limits_{N \to \infty} E_N(\delta_i) = p > 0 & \text{i.e. non-negligible B-selection} \end{cases} \tag{17}$$

The NPA assumption (17) encompasses both the SP and QR approach. For the SP approach, taking the conditional expectation of y_i's conditional on the δ_i's yields

$$E\big(Cov_N(\delta_i, y_i) | \delta_U\big) = \frac{1}{N} \sum_{i \in U} \delta_i \mu_i - \big(\frac{1}{N} \sum_{i \in U} \delta_i\big)\big(\frac{1}{N} \sum_{i \in U} \mu_i\big) \to 0$$

given NPA non-informative B-selection, where $\sum_{i \in U} \delta_i / N > 0$ given non-negligible B-selection in addition. Under this condition, the B-sample expansion estimator (1) is asymptotically prediction unbiased from the SP perspective. For the QR approach, taking the expectation of δ_i's with the y_i's being constants yields

$$\begin{cases} E\big(Cov_N(\delta_i, y_i); y_U\big) = \frac{1}{N} \sum\limits_{i \in U} p_i y_i - \big(\frac{1}{N} \sum\limits_{i \in U} p_i\big)\big(\frac{1}{N} \sum\limits_{i \in U} y_i\big) \to 0 \\ E\big(E_N(\delta_i)\big) = \sum_{i \in U} p_i / N \to p > 0 \end{cases}$$

In particular, the NPA assumption (17) allows for $0 \leq p_i \leq 1$, so that the B-sample expansion estimator (1) remains consistent from the QR perspective, even in the presence of under-coverage of the units with $p_i = 0$ or non-representative units with $p_i = 1$.

Example: Let $U = U_1 \cup U_0$ be a partition. Let U_1 be of size N_1, where $p_i \equiv 1$ for $i \in U_1$; let U_0 be of size N_0, where $p_i \equiv 0$ for $i \in U_0$. Despite under-coverage of $B \equiv U_1$, the first NPA condition implies $\bar{y}_B - \bar{Y} \to 0$, given the second condition $N_1/N \to p > 0$.

3.2 Post-stratification estimator

Consider post-stratification by x_i, for $i \in U$. Provided the assumption (17) holds within each post-stratum, the B-sample post-stratification estimator is asymptotically unbiased from both the SP and QR perspective. Below we consider the QR approach. The SP approach is a special case of the calibration estimator discussed in Section 3.3.

Consider first the hypothetical estimator with known $p_x = \sum_{i \in U_x} p_i / N_x$:

$$\widetilde{Y} = \sum_x \sum_{i \in U_x} \delta_i y_i / p_x$$

To fix the idea for variance estimation, suppose independent Bernoulli distribution of δ_i with probability p_i, where $0 \leq p_i \leq 1$. The variance of $\widetilde{Y}$ is then given by

$$V(\widetilde{Y}) = \sum_x \sum_{i \in U_x} p_i y_i^2 / p_x^2 - \sum_x \sum_{i \in U_x} p_i^2 y_i^2 / p_x^2$$

An unbiased estimator of the first term of the variance, denoted by τ_1 is given by

$$\hat{\tau}_1 = \sum_x \sum_{i \in U_x} \delta_i y_i^2 / p_x^2 = \sum_x p_x^{-2} \sum_{i \in B_x} y_i^2$$

where $B_x = B \cap U_x$. An unbiased estimator of the second term, denoted by τ_2 is given by

$$\hat{\tau}_2 = \sum_x p_x^{-2} \sum_{i \in U_x} \delta_i p_i y_i^2 = \sum_x p_x^{-1} \sum_{i \in U_x} \delta_i y_i^2 = \sum_x p_x^{-1} \sum_{i \in B_x} y_i^2$$

where the second equality follows given the additional QR$_x$ assumption, i.e. $p_i = p_x$ for $i \in U_x$. Putting $\hat{\tau}_1$ and $\hat{\tau}_2$ together, we obtain

$$\widehat{V}(\widetilde{Y}) = \sum_x \big(p_x^{-1} - 1\big) p_x^{-1} \sum_{i \in B_x} y_i^2$$

Now, the post-stratification estimator, denoted by $\widehat{Y}$, is obtained from $\widetilde{Y}$ on replacing p_x by $\hat{p}_x = n_{xB}/N_x$, where n_{xB} is the observed size of B_x and N_x is the known post-stratum population size. Expanding $\hat{p}_x$ around p_x (i.e., linearisation) would yield an asymptotically valid estimator of the unconditional variance of $\widehat{Y}$.

3.3 Calibration estimator

The post-stratification estimator is infeasible, in cases when the B-sample contains empty cells, or when the population size N_x is not all known. Let

$$t_i = (t_{1i}, t_{2i}, ...t_{Ki})^\top = \big(t_1(x_i), t_2(x_i), ...t_K(x_i)\big)^\top = t(x_i)$$

be a vector of many-to-one mappings of x_i, such that the population total $T = \sum_{i \in U} t_i$ is known, and the sample total $t = \sum_{i \in B} t_i$ has only non-zero components.

As discussed for the calibration estimator in Section 2, generally one is not able to set the initial weight to be the inverse of B-sample inclusion probability in practice. Suppose one simply starts with initial equal weights $a_i = N/n_B$ for all $i \in B$. The linear calibration estimator (Deville and Särndal, 1992) is given by

$$\widehat{Y} = \sum_{i \in B} w_i y_i$$

where the weights $\{w_i; i \in B\}$ minimise the distance to $\{a_i; i \in B\}$ as measured by

$$\sum_{i \in B} (w_i - N/n_B)^2 = \sum_t \big(\sum_{i \in B_t} w_i^2 - 2(N/n_B) \sum_{i \in B_t} w_i + n_{tB}(N/n_{tB})^2 \big)$$

subjected to the constraints $\sum_{i \in B} w_i t_i = T$, where $B_t = \{i; t_i = t, i \in B\}$ and $n_{tB} > 0$. It follows that $w_i = w_t$, for $i \in B_t$, since the only thing that matters to the calibration constraints is $\sum_{i \in B_t} w_i$ now that $t_i = t$ for $i \in B_t$ and, given whatever $\sum_{i \in B_t} w_i$, the term $\sum_{i \in B_t} w_i^2$ is minimised at $w_i = w_t$ for $i \in B_t$.

As the first validity condition for $\widehat{Y}$, suppose there exists a vector $\beta_{K \times 1}$, such that

$$\sum_{i \in U_t} \epsilon_i / N_t \to 0 \tag{18}$$

for each t-value, as $N \to \infty$, where $\epsilon_i = y_i - t_i^\top \beta$, and N_t is the population size of $U_t = \{i; t_i = t, i \in U\}$. The condition (18) is analogous to the SP$_x$ assumption (6), where the covariate x_i is replaced by t_i here. Moreover, it relaxes the model (6) of the conditional mean, allowing for potential heterogeneous mean similar to (15). Now that $\sum_{i \in B} w_i t_i = T$, we have

$$\widehat{Y} - Y = \sum_{i \in B} w_i (t_i^\top \beta + \epsilon_i) - \sum_{i \in U} t_i^\top (\beta + \epsilon_i) = \sum_{i \in B} w_i \epsilon_i - \sum_{i \in U} \epsilon_i$$

Given (18), $\sum_{i \in U} \epsilon_i / N \to 0$ as $N \to \infty$. Moreover, we have

$$\frac{1}{N} \sum_{i \in B} w_i \epsilon_i = \sum_t \frac{w_t}{N} \sum_{i \in U_t} \delta_i \epsilon_i = \sum_t w_t \frac{N_t}{N} \big(Cov_{N_t}(\delta_i, \epsilon_i) + (\frac{1}{N_t} \sum_{i \in U_t} \delta_i)(\frac{1}{N_t} \sum_{i \in U_t} \epsilon_i) \big) \to 0$$

as $N \to \infty$, given

$$\begin{cases} Cov_{N_t}(\delta_i, \epsilon_i) \to 0 \\ E_{N_t}(\delta_i) = \sum_{i \in U_t} \delta_i / N_t \to p_t > 0 \end{cases} \tag{19}$$

for any given t, which is an adaption of the NPA non-informativeness assumption (17) to the present setting. It follows that the two assumptions (18) and (19) are the key validity conditions for the calibration estimator to be consistent for Y.

For variance estimation, suppose again independent Bernoulli distribution of δ_i with probability p_i, where $0 \leq p_i \leq 1$. An approximate variance estimator for the calibration estimator $\widehat{Y}$ can then be given as

$$\widehat{V}(\widehat{Y}) = \sum_t \big(\hat{p}_t^{-1} - 1\big)\hat{p}_t^{-1} \sum_{i \in B_t} (y_i - t_i^\top \hat{\beta})^2$$

where $\hat{p}_t = n_{tB}/N_t$, and $\hat{\beta} = \big(\sum_{i \in B} w_i t_i t_i^\top\big)^{-1}\big(\sum_{i \in B} w_i t_i y_i\big)$.

3.4 Validation of non-informative B-sample selection

Of the validity conditions discussed above, the critical assumption is non-informative B-sample selection, which can be stated in various forms. For instance, given the non-informativeness assumption (17), an additional assumption like (18) can in principle to validated empirically. However, the non-informativeness condition may not hold exactly, and it is generally impossible to verify only based on the data used for the estimation. Below we discuss the issue in more detail.

Consider first the propensity model $p_i = p(x_i; \eta)$ under the QR approach. Suppose known x_U to avoid additional complications otherwise, the census score equation is

$$\sum_x \frac{\partial p(x;\eta)}{\partial \eta} \left[\frac{n_{xB}}{p(x;\eta)} - \frac{N_x - n_{xB}}{1 - p(x;\eta)} \right] = 0$$

which is always satisfied by $p(x; \hat{\eta}) = n_{xB}/N_x$, i.e. the saturated model. Insofar as a non-saturated model of $p(x_i; \eta)$ does not fit perfectly to the data, one can always attribute its cause to the non-saturated functional form of $p(x_i; \eta)$, instead of rejecting the assumption that the set of covariates x_i "fully govern the sampling mechanism". In this sense the validity of the latter assumption cannot be refuted empirically.

Next, for the SP approach, where both δ_i and y_i are treated as random, assume the B-sample inclusion probability p_i depend on x_i, where x_i is known for $i \in U$ to avoid extra complications. For goodness-of-fit checks, let z_i be a known covariate, which is distinct from x_i. We have

$$\begin{cases} E(z_B) = \sum_{i \in U} p_i z_i = \sum_x p(x;\eta) \sum_{i \in U_x} z_i = \sum_x p(x;\eta) N_z \bar{Z}_x \\ Z = E(\sum_{i \in U} \delta_i z_i / p_i) = E[\sum_x n_{xB} \bar{z}_{xB} / p(x;\eta)] \end{cases}$$

where $\bar{Z}_x = \sum_{i \in U_x} z_i / N_x$ and $\bar{z}_{xB} = \sum_{i \in B_x} z_i / n_{xB}$. The two observed checks are

$$\begin{cases} z_B \equiv \sum_x n_{xB} \bar{z}_{xB} = \sum_x \hat{p}_x N_x \bar{Z}_x \\ Z = \sum_x n_{xB} \bar{z}_{xB} / \hat{p}_x \end{cases} \overset{\text{if } z_i \equiv 1}{\Longrightarrow} \begin{cases} \sum\limits_{i \in U} \hat{p}_i = n_B \\ \sum\limits_{i \in B} 1/\hat{p}_i = N \end{cases}$$

Setting $\hat{p}_x = n_{xB}/N_x$, which fits the assumption $p_i = p(x_i; \lambda)$, both the two checks are satisfied given $\bar{Z}_x = \bar{z}_{xB}$, i.e. the B-sample expansion estimate of Z_x is perfect for all x. This would suggest that the NPA assumption (17) holds for z_i given x_i, and may be considered to support the plausibility of the NPA assumption (17) for y_i given x_i, provided *z_i is known to be correlated with y_i*, but not otherwise. However, in situations where such a covariate z_i is available, it seems natural that it should be used in the estimation of Y to start with. The two checks amounts then to the case of $z_i \equiv 1$, and are satisfied trivially by setting $\hat{p}_x = n_{xB}/N_x$. Thus, one is faced with a dilemma, where building the best model for estimation would at the same time reduce the ability to verify it.

4 Using additional probability sample of outcomes

So far we have only considered the situations, where the outcome values of interest are *only* observed in the non-probability sample B. Obviously, the situation changes completely, given *in addition* a probability sample of outcomes. Below we discuss shortly two different approaches to inference in the absence of any relevant covariates. The ideas remain the same in situations with additional covariates.

The first approach aims at consistent estimation combing the two samples, as e.g. discussed in Tam and Kim (2018a, 2018b), where the probability sample is taken from the whole population and overlaps with the B-sample. These authors also discussed additional issues such as measurement errors or nonresponse. Here we discuss the situation where the probability sample is taken from the B-sample complement population. Given the non-probability sample observations y_B, one may treat (B, y_B) as fixed, and select a second supplementary sample from the rest of the population, denoted by $S \subset U \setminus B$. Given the S-sample observations of the outcome, denoted by y_S, it is straightforward to obtain a test for $H_0 : \bar{Y} = \bar{y}_B$ vs. $H_1 : \bar{Y} \neq \bar{y}_B$, given as

$$D = (\bar{y}_B - \widehat{\bar{Y}_B^c})^2 / \widehat{V}(\widehat{\bar{Y}_B^c}) \;\sim\; \chi_1^2$$

where $\widehat{\bar{Y}_B^c}$ is an S-sample estimator of the population mean outside of the B-sample, i.e.

$$\bar{Y}_B^c = \sum_{i \in U \setminus B} y_i / (N - n_B)$$

and $\widehat{V}(\widehat{\bar{Y}_B^c})$ is the associated variance estimator. If H_0 is not rejected, then there is the possibility of using $\bar{y}_B$ as an estimate on its own, without regular concurrent surveys in future. This would achieve the greatest cost savings. To this end, one may consider S as a particular form of audit sampling, whose aim is to validate the big-data estimate $\bar{y}_B$ and to provide a meaningful accuracy measure that can accommodate its potential bias. Zhang (2019) develops an approach to audit sampling inference for big data statistics.

Let $W_B = n_B/N$. A consistent estimator of $\bar{Y}$ using both samples is given by

$$\widehat{\bar{Y}}_S = W_B \bar{y}_B + (1 - W_B)\bar{y}_w \qquad \text{and} \qquad \bar{y}_w = \frac{\sum_{i \in S} y_i/\pi_i}{\sum_{i \in S} 1/\pi_i}$$

where π_i is the S-sample inclusion probability, and the validity of $\widehat{\bar{Y}}_S$ now derives from probability sampling of S, regardless of how the B-sample is generated. The relative efficiency (RE) against the setting without the B-sample can be given by

$$\text{RE} = \big[(1 - W_B)^2 V(\bar{y}_w)\big] / V(\widehat{\bar{Y}'})$$

where $\widehat{\bar{Y}'}$ is a hypothetical probability sample from the whole population U, which has the same sample size and the same sampling design as S. One may refer to this as the *split-population* approach to inference, which is an age-old idea for combining survey sampling with administrative data. The efficiency gain would be substantial provided the B-sample is large. In fact, the larger the B-sample, the greater is the efficiency gain.

Under the second approach to inference, consider a *composite* estimator given by

$$\widehat{\bar{Y}}_C = \gamma \bar{y}_B + (1 - \gamma)\bar{y}_w$$

where γ is the composition weight, for $W_B \leq \gamma \leq 1$. Notice that when $\gamma = W_B$, the composite estimator is just the split-population estimator $\widehat{\bar{Y}}_S$ above, which is consistent for $\bar{Y}$. As γ increases from W_B towards one, one risks introducing greater bias, insofar as $\bar{y}_B \neq \bar{Y}$. However, the composite estimator may yield a smaller mean squared error (MSE) of estimation, provided this is desirable. One is then essentially trading the increasing bias $(\gamma - W_B)(\bar{y}_B - \bar{Y}_B^c)$ against the decreasing stand error $(1 - \gamma)\text{SE}(\bar{y}_w)$, as γ increases. The composite estimator that achieves the minimum MSE is given by

$$\gamma = \frac{V(\bar{y}_w) + W_B(\bar{y}_B - \bar{Y}_B^c)^2}{V(\bar{y}_w) + (\bar{y}_B - \bar{Y}_B^c)^2}$$

Estimating $\bar{Y}_B^c$ by $\bar{y}_w$ in application, one can use

$$\hat{\gamma} = \min(W_B + (1 - W_B)\widehat{V}(\bar{y}_w)/(\bar{y}_B - \bar{y}_w)^2, 1)$$

The validity of the composite approach derives from probability sampling of S, regardless of how the B-sample is generated. Again, the bigger the B-sample, the better it is.

5 Summary

All the estimators from non-probability sample observations reviewed in Section 2 are model-based, whether the modelling is carried out under the SP or QR approach. Two features regarding the model covariate x_i, for $i \in U$, are worth recapitulating:

- compared to the situation with known x_U, making use of an additional probability sample x_S entails a loss of efficiency, as can be expected;
- the availability of an additional probability sample without the outcome variable is not a principal advantage, since it does not simplify the validity conditions compared to the situation where x_U is known, but it does resolve the practical difficulty when x_U is unavailable yet some functions of x_U are needed for descriptive inference.

The situation changes completely, given in addition a probability sample of outcomes. The probability sample then enables valid descriptive inference in combination with the non-probability probability sample. Depending on the circumstances, the probability sample can either be selected from the whole population, or just the rest population outside the non-probability sample.

Finally, in situations where the non-probability sample is large, the cost savings will be the greatest if it can replace regular survey sampling altogether. Use of an additional probability audit sample is needed to validate the non-probability sample estimate, in spite of possible failure of its underlying model assumptions, and to provide a meaningful accuracy measure that can accommodate its potential bias.

References

[1] Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. and Tourangeau, R. (2013). *Report of the AAPOR Task Force on Non-probability Sampling.* Technical report, American Association for Public Opinion Research, Deerfield, IL.

[2] Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, **4**, 251-260.

[3] Chen, J.H. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, **16**, 113-131.

[4] Deville J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, **87**, 376-382.

[5] Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, **32**, 249-264.

[6] Kang, J.D.Y. and Schafer, J.L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, **22**, 523-539.

[7] Kim, J.-K. and Wang, Z. (2018). Sampling techniques for big data analysis in finite population inference. arXiv:1801.09728v1

[8] Kim, J.-K. and Rao, J.N.K. (2018). Data Integration for Big Data Analysis in Finite Population Inference. *Talk presented at SSC2018, Montreal.*

[9] Kim, J.K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica* **24**, 375-394.

[10] Little, R. J. A. (1982) Models for nonresponse in sample surveys. *Journal of the American Statistisical Association*, **77**, 237-250.

[11] Meng, X.L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and 2016 US presidential election. *Annals of Applied Statistics*, **12**, 685-726.

[12] Oh. H. L. and F. J. Scheuren (1983). Weighting adjustments for unit non-response. In W. G. Madow, I. Olkin and D. B. Rubin (Eds.), *Incomplete data in sample surveys (Vol. 2): Theory and bibliographies*, pp. 143-184. Academic Press (New York; London).

[13] Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, **8**, 1087-1114.

[14] Pfeffermann, D. (2017). Bayes-based Non-Bayesian Inference on Finite Populations from Non-representative Samples. *Calcutta Statistical Association Bulletin*, **69**, 1-29. DOI:10.1177/0008068317696546

[15] Rao, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhya*, **28**, 47-60.

[16] Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

[17] Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficient when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846-866.

[18] Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41-55.

[19] Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, **57**, 377-387.

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

[21] Smith, T.M.F. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society, Series A*, **146**, 394– 403.

[22] Tam, S.-M. and J.-K. Kim (2018a). Big Data ethics and selection-bias: An official statistician's perspective. *Statistical Journal of the IAOS,* **34***: 577-588.* DOI: `10.3233/SJI-170395`

[23] Tam, S.-M. and J.-K. Kim (2018b). Mining Big Data for Finite Population Inference. *Talk presented at BigSurv18, Barcelona.*

[24] Yang, S. and J.-K. Kim (2018). Integration of survey data and big observational data for finite population inference using mass imputation. `https://arxiv.org/abs/1807.02817v1`

[25] Zhang, L.-C. (2019). Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big data statistics. `https://arxiv.org/abs/1906.11208`