Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Comments on “Statistical inference with non-probability survey samples” – Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples

by Xiao-Li Meng

Release date: December 15, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistical Information Service	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Comments on "Statistical inference with non-probability survey samples" – Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples

Xiao-Li Meng[1]

Abstract

Non-probability samples are deprived of the powerful *design probability* for randomization-based inference. This deprivation, however, encourages us to take advantage of a natural *divine probability* that comes with any finite population. A key metric from this perspective is the *data defect correlation (ddc)*, which is the model-free finite-population correlation between the individual's sample inclusion indicator and the individual's attribute being sampled. A data generating mechanism is equivalent to a probability sampling, in terms of design effect, if and only if its corresponding *ddc* is of $N^{-1/2}$ (stochastic) order, where N is the population size (Meng, 2018). Consequently, existing valid linear estimation methods for non-probability samples can be recast as various strategies to miniaturize the *ddc* down to the $N^{-1/2}$ order. The quasi design-based methods accomplish this task by diminishing the variability among the N inclusion propensities via weighting. The super-population model-based approach achieves the same goal through reducing the variability of the N individual attributes by replacing them with their residuals from a regression model. The doubly robust estimators enjoy their celebrated property because a correlation is zero whenever one of the variables being correlated is constant, regardless of which one. Understanding the commonality of these methods through *ddc* also helps us see clearly the possibility of "double-plus robustness": a valid estimation without relying on the full validity of either the regression model or the estimated inclusion propensity, neither of which is guaranteed because both rely on *device probability*. The insight generated by *ddc* also suggests *counterbalancing sub-sampling*, a strategy aimed at creating a miniature of the population out of a non-probability sample, and with favorable quality-quantity trade-off because mean-squared errors are much more sensitive to *ddc* than to the sample size, especially for large populations.

Key Words: Data defect index; Design probability; Divine probability; Device probability; Design-based inference; Model-assisted survey estimators; Non-response bias.

1. Distinguish among design, divine, and device probabilities

1.1 What can statistics/statisticians say about non-probability samples?

Dealing with non-probability samples is a delicate business, especially for statisticians. Those who believe statistics is all about probabilistic reasoning and inference may question if statistics has anything useful to offer to the non-probabilistic world. Whereas such questioning may reflect the inquirers' ignorance about or even hostility towards statistics, taking the question conceptually, it deserves statisticians' introspection and extrospection. What kind of probabilities are we referring to when the sample is non-probabilistic? The entire probabilistic sampling theory and methods are built upon the randomness introduced by powerful sampling mechanisms, which then yields the beautiful designed-based inferential framework without having to *conceive* anything else is random (Kish, 1965; Wu and Thompson, 2020; Lohr, 2021). When that power – and beauty – is taken away from us, what's left for statisticians?

[1] Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, MA 02138. E-mail: meng@stat.harvard.edu.

A philosophical answer by some statisticians would be to dismiss the question altogether by declaring that there is no such thing as probability sample in real life. (I was reminded by Andrew Gelman about this sentiment when I sought his comments on this discussion article. See https://statmodeling.stat.columbia.edu/2014/08/06 for a related discussion.) By the time the data arrive at our desk or disk, even the most carefully designed probability sampling scheme would be compromised by the imperfections in execution, from (uncontrollable) defects in sampling frames to non-responses at various stages and to measurement errors in the responses. In this sense, the notion of probability sample is always a theoretical one, much like efficient market theory in economics, which offers a mathematically elegant framework for idealization and for approximations, but should never be taken literally (e.g., Lo, 2017).

The timely article by Professor Changbao Wu (Wu, 2022) provides a more practical answer, by showcasing how statisticians have dealt with non-probability samples in the long literature of sample surveys and (of course) observational studies, especially for causal inference; see Elliott and Valliant (2017) and Zhang (2019) for two complementary overviews addressing the same challenge. To better understand how probability theory is useful for non-probability samples, it is important to recognize (at least) three types of probabilistic constructs for statistical inference, as listed in Section 1.2. Non-probability samples take away only one of the three, and as a result, they typically force a stronger reliance on the other two.

With these conceptual issues clarified, the rest sections discuss a unified strategy for dealing with non-probability samples. Section 2 reviews a fundamental identity for estimation error, which has led to the construction of data defect correlation (Meng, 2018). Section 3 then discusses how this construct suggests the unified strategy. Section 4 demonstrates the strategy respectively for the qp and ξp settings in Wu (2022). Section 5 then applies the strategy to the two settings simultaneously to reveal an immediate insight into the celebrated double robustness, as reviewed in Wu (2022). Inspired by the same construct, Section 6 explores *counterbalancing sampling* as an alternative strategy to weighting. Section 7 concludes with a general call to treat probability sampling theory as an aspiration instead of the centerpiece of survey and sampling research.

1.2 A trio of probability constructs

The first of the three named constructs below, design probability, is self-explanatory. It is at the heart of sampling theory and reified by practical implementation, however imperfect the implementation might be. The distinction between the next two, divine probability and device probability, may be more nuanced especially at practical levels. But their conceptual differences are no less important than distinguishing between an estimand and an estimator. Fittingly, the data recording or inclusion indicator, a key quantity in modeling non-probability samples, provides a concrete illustration of all three probabilistic constructs; see the leading paragraph of Section 4.

Design Probability. A paramount concept and tool for statistics – and for general science – is randomized replications (Craiu, Gong and Meng, 2022). By designing and executing a probabilistic mechanism to generate randomized replications, we create probabilistic data that can be used directly for making verifiable inferential statements. Besides probabilistic sampling in surveys, randomization in clinical trials, bootstraps for assessing variability, permutation tests for hypothesis testing, and Monte Carlo simulations for computing are all examples of statistical methods that are built on design probability. Non-probability samples, by definition, do not come with such design probability, at least not an identified one. Hence, the phrase non-probability samples should be understood as a short hand for “samples without an identified design probability construct”.

It is worth to remind ourselves, however, that there is a potential for design probabilities to come back in a substantial way especially for large non-probability data sets, such as administrative data, due to the adoption of differential privacy (Dwork, 2008), for example by US Census Bureau (see the editorial by Gong, Groshen and Vadhan, 2022, and the special issue in *Harvard Data Science Review* it introduces). Differential privacy methods inject well-designed random noise into data for the purpose of protecting data privacy while not unduly sacrificing data utility. Like the design probability used for probabilistic sampling, the fact that the noise-injecting mechanism is designed by the data curator, and is made publicly known, renders the transparency that is critical for valid statistical inference by the data user (Gong, 2022). The area of how to properly analyze non-probability data with differential privacy protection is wide open. Even more so is the fascinating area of how to take into account the existing defects in non-probability data when designing probabilistic protection mechanisms for data privacy to avoid adding unnecessary noise. Readers who are interested in forming a big picture of the statistical issues involved in data privacy should consult the excellent overview article by Slavkovic and Seeman (2022) on the general area of “statistical data privacy”.

Divine Probability. In the absence of design probability for randomization-based inference, in order to conduct a (conventional) statistical inference, we typically conceptualize that the data at hand is a realization of a generative probabilistic mechanism given by nature or God. (I learned about the term “God’s model” during my PhD training, which I took as an expression for faith or something beyond human control, rather than reflecting one’s religious belief. The phrase “divine” is adopted here with a similar connotation.) We do so regardless of whether we believe or not that the world is intrinsically deterministic or stochastic (e.g., see David Peat, 2002; Li and Meng, 2021). We need to assume this divine probability primarily because of the restrictive nature of the probabilistic framework to which we are so accustomed. For example, in order to invoke the assumption of missing at random, we need to conjure a probabilistic mechanism under which the concept “missing at random” (Rubin, 1976) can be formalized. As Elliott and Valliant (2017) emphasized, the quasi-randomization approach, which corresponds to the qp framework of Wu (2022), “assumes that the nonprobability sample actually does have a probability sampling mechanism, albeit one with probabilities that have to be estimated under identifying assumptions". That is, we replace the design probability by a divine probability that we have faith for its existence, which then typically is treated as the "truth" or at least as an estimand.

Conceptually, therefore, we need to recognize that the assumption of any particular kind of divine probability is not innocent, as otherwise we will not need to rely on our faith to proceed. Nor is it always necessary. Any finite population provides a natural histogram for any quantifiable attributes or a contingency table for any categorizable attributes of its constituents, and hence it induces a divine probability without referencing any kind of randomness, conceptualized or realized, *if our inferential target is the finite population itself* (not a super-population that generates it, for example). The empirical likelihood approach takes advantage of this natural probability framework, which also turns out to be fundamental for quantifying data quality via data defect correlation (see Meng, 2018). The same emphasis was made by Zhang (2019), whose unified criterion was based on the same identity for building data defect correlation; see Section 2 below.

Device Probability. By far, most probabilities used in statistical modeling are devices for expressing our belief, prior knowledge, assumptions, idealizations, compromises, or even desperation (e.g., imposing a prior distribution to ensure identifiability since nothing else works). Whereas modeling reality has always been a key emphasis in the statistical literature, we inevitably must make a variety of simplifications, approximations, and some times deliberate distortions in order to deal with practical constraints (e.g., the use of variational inference for computational efficiency; see Blei, Kucukelbir and McAuliffe (2017)). Consequently, many of these device probabilities do not come with a requirement of being realizable, or even coherent mathematically (e.g., the employment of incompatible conditional probability distributions for multiple chain imputation; see Van Buuren and Oudshoorn (1999)). Nor are they easy or even possible to be validated, as Zhang (2019) investigated and argued in the context of non-probability sampling, especially with the superpopulation modeling approach, which corresponds to the ξp framework of Wu (2022). Nevertheless, device probabilities are the workhorse for statistical inferences. Both quasi-randomization approach and super-population modeling rely on such device probabilities to operate, as shown in Wu (2022) and further discussed in Sections 4-5 below. The lack of design probability can only encourage more device probabilities to make headway. To paraphrase Box's famous quote "all models are wrong, but some are useful", all device probabilities are problematic, but some are problem-solving.

1.3 Let's reduce "Garbage in, package out"

In a nutshell, probabilistic constructs are more needed for non-probability samples than probability ones precisely because of the deprivation of the design probability. Therefore, dealing with non-probability samples is not a new challenge for statisticians. If anything is new, it is the availability of massive amounts of large and non-probabilistic data sets, such as administrative data and social media data, and the accelerated need to combine multiple sources of data, most of which inherently are non-probabilistic because they are not collected for statistical inference purposes (e.g., Lohr and Rao, 2006; Meng, 2014; Buelens, Burger and van den Brakel, 2018; Beaumont and Rao, 2021). Contrary to common belief, the large sizes of “big data” can make our inference much worse, because of the “big data paradox” (Meng, 2018; Msaouel, 2022) when we fail to take into account the data quality in assessing the errors and uncertainties in our analyses; see Section 6.1.

It is therefore becoming more pressing than ever to greatly increase the general awareness of, and literacy about, the critical importance of data quality, and how we can use statistical methods and theories to help to reduce the data defect. The central concern here goes beyond the common warning about “garbage in, garbage out” – if something is recognized as garbage, it would likely be treated as such (likely, but not always, because as Andrew Gelman reminded me that “many researchers have a strong belief in *procedure* rather than *measurement*, and for these people the most important thing is to follow the rules, not to look at where their data came from”). The goal is to prevent “garbage in, package out” (Meng, 2021), where low quality data are auto-processed by generic procedures to create a cosmetically attractive “AI” package and sold to uninformed consumers or worse, to those who seek “data evidence” to mislead or disinform. Properly handling non-probability samples obviously does not resolve all the data quality issues, but it goes a very long way in addressing an increasingly common and detrimental problem of lack of data quality control in data science.

I therefore thank Professor Changbao Wu for a well timed and designed in-depth tour of “the must-sees” of the large sausage-making factory for processing non-probability samples. It adds considerably more detailed and nuanced exhibitions to the general tour by Elliott and Valliant (2017), which includes excellent illustrations on many forms and shapes of non-probability samples as well as their harms. It also showcases theoretical and methodological milestones for us to better appreciate the millstones displayed in the intellectual tour by Zhang (2019), which challenges statisticians and data scientists in general to understand better the quality, or rather the lack thereof, of the products we produce and promote. Together, this trio of overview articles form an informative tour for anyone who wants to join the force to address the ever-increasing challenges of non-probability data. Perhaps the best tour sequence starts with Elliott and Valliant (2017) to form a general picture, with Wu (2022)’s as the main exhibition of methodologies, and ends with Zhang (2019) to generate deep reflections on some specific challenges. For additional common methods for dealing with non-probability samples, such as multilevel modeling and poststratification, readers are referred to Gelman (2007), Wang, Rothschild, Goel and Gelman (2015) and Liu, Gelman and Chen (2021).

As a researcher and educator, I have been beating similar drums but often frustrated by the lack of time or energy to engage deeply. I am therefore particularly grateful to Editor Jean-François Beaumont for inviting me to help to ensure Professor Wu’s messages are loud and clear: data cannot be processed as if they were representative unless the observed data are genuinely probability samples (which is extremely rare); many remedies have been proposed and tried, but many more need to be developed and evaluated. Among them, the concept of data defect correlation is a promising general metric to be explored and expanded, as demonstrated below.

2. A finite-population deterministic identity for actual error

To demonstrate the fruitfulness of the finite-population framework, consider the estimation of the population mean, denoted by $\bar{G}$, of $\{G_i = G(X_i): i \in \mathcal{N}\}$, where $\mathcal{N} = \{1, \ldots, N\}$ indexes a finite population, and the X_i's are data collected on individual i. For each i, let $R_i = 1$ if G_i (or rather X_i) is recorded in our sample, and $R_i = 0$ otherwise; hence the sample size is $n_R = \sum_{i=1}^N R_i$. We stress that this is an all-encompassing indicator, which can (and should) be decomposed into $R_i = r_i^{(1)}, \ldots, r_i^{(J)}$, when the data collection consists of J stages (e.g., $r_i^{(1)}$ indicates whether or not the i^{th} individual is sampled, and $r_i^{(2)}$ whether the individual responded or not once sampled).

Let $\{W_i, i \in S\}$ be a set of weights to be determined, where the index set $S = \{i : R_i = 1\}$, such that $\sum_{i \in S} W_i > 0$. Let $\bar{G}_W$ be the weighted sample average, expressible in three ways:

$$\bar{G}_W = \frac{\sum_{i \in S} W_i G_i}{\sum_{i \in S} W_i} = \frac{\sum_{i=1}^N R_i W_i G_i}{\sum_{i=1}^N R_i W_i} = \frac{\mathrm{E}_I(\tilde{R}_I G_I)}{\mathrm{E}_I(\tilde{R}_I)}, \tag{2.1}$$

where $\tilde{R}_I = R_I W_I$, and E_I is taken with respect to the uniform distribution over the index set $\mathcal{N}$. The first expression in (2.1) simply defines a weighted sample average. With the help of R_i, the second expression turns the sample averages into finite-population averages. This trivial re-expression is fundamental because it explicates the role of R_i in influencing the behavior of $\bar{G}_W$ as an estimator of $\bar{G}$. The third expression reveals a divine probability through I, the finite-population index (FPI) variable, by utilizing the fact that averaging is the same as taking expectation over a uniformly distributed random index I. All finite-population moments then can be expressed via E_I.

In particular, we can express the actual error of $\bar{G}_W$ via the following identity, where the first expression can be traced back to Hartley and Ross (1954), who used it to express biases in ratio estimators. The second expression was given in Meng (2018) with a slightly different (but equivalent) expression:

$$\bar{G}_W - \bar{G} = \frac{\mathrm{Cov}_I(\tilde{R}_I, G_I)}{\mathrm{E}_I[\tilde{R}_I]} = \rho_{\tilde{R},G} \times \sqrt{\frac{N - n_W}{n_W}} \times \sigma_G. \tag{2.2}$$

Here $\rho_{\tilde{R},G} = \mathrm{Corr}_I(\tilde{R}_I, G_I)$ is the *finite-population correlation* between $\tilde{R}_I$ and G_I, σ_G^2 is the finite-population variance of G_I, and n_W is the effective sample size due to using weights (Kish, 1965)

$$n_W = \frac{n_R}{1 + \mathrm{CV}_W^2}, \tag{2.3}$$

with CV_W being the coefficient of variation (i.e., standard deviation/mean) of $\{W_i, i \in S\}$.

The expression (2.2) is an algebraic identity because it holds for any instances of $\{(G_i, R_i W_i), i \in \mathcal{N}\}$. Hence no model assumptions are imposed, not even the assumption that R (or any quantity) is random, echoing the comment by Mary Thompson, as quoted in Wu (2022), that "the sample inclusion indicator R is a random variable is itself an assumption". The only requirement is that the recorded G_i is unchanged from the G_i's in the target population. (But note this requirement has two components: (1) there is no over-coverage, that is, everyone in the sample belongs to the target population, e.g., no non-eligible voters are surveyed when the target population is eligible voters, and (2) there is no measurement error; extensions to the cases with measurement errors are available, but not pursued in this article.) When we use equal weights, the three factors on the right-hand side of (2.2) reflect respectively (from left to right) data defect, data sparsity, and problem difficulty, as detailed in Meng (2018) and further illustrated in Bradley, Kuriwaki, Isakov, Sejdinovic, Meng and Flaxman (2021) in the context of COVID-19 vaccination surveys.

In particular, when all weights are equal, $\rho_{\tilde{R},G}$ is termed as *data defect correlation* (*ddc*) in Meng (2018) because it measures the lack of representativeness of the sample via capturing the dependence of inclusion/recording indicator on the attributes – the higher the dependence, the more biased the sample average becomes for estimating population averages. With the basic strategies of probabilistic sampling or inverse probability weighting, *ddc* will be zero on average because $\mathrm{E}(W_i R_i) = 1,$ and it is of $O_p(N^{-1/2})$ order because it is essentially an average of N independent terms (Meng, 2018). Our general goal here therefore is to bring down *ddc* to $O_p(N^{-1/2})$ for non-probability samples, which we shall refer to as "miniaturizing *ddc*" because $N^{-1/2}$ is typically a minuscule number in practice.

When we use weights, the first term $\rho_{\tilde{R},G}$ captures the data defect that still exists after the weighting adjustment, since no weights are perfect in practice. Identity (2.2) shows the impact of the weights on both data quality and data quantity. The impact on the *nominal* effective sample size n_W is never positive because $n_W \leq n_R$ as seen in (2.3). Incidentally, the exactness of (2.3) reveals that this well-known expression is in fact not an approximation (which is often attributed to Kish (1965)), but an exact formula for the reduction of the sample size due to weighting *if the weighting had no impact on ddc*. However, weighting can have a major positive impact on reducing the overall error by judiciously choosing weights to significantly decrease *ddc*, though apparently at the price of $n_W < n_R$. Of course, this is exactly the aim of the quasi-randomization framework, as discussed below. Most importantly, however, (2.2) leads to a unified insight about the variety of methods reviewed in Wu (2022), including an intuitive explanation of the doubly robust property, which has been receiving increased attention for integrating data sources including both probability and non-probability samples (e.g., Yang, Kim and Song, 2020).

Indeed, Zhang (2019, Section 3.1) used the first expression in (2.2) to define a unified non-parametric asymptotic (NPA) non-informativeness assumption, which requires that the numerator $\mathrm{Cov}_I(\tilde{R}_I, G_I)$ goes to zero, while keeping the denominator $\mathrm{E}_I[\tilde{R}_I]$ positive, as $N \to \infty$. This unification permits Zhang (2019) to evaluate the quasi-randomization approach and regression modeling via a common criterion. The *ddc* framework echoes this unification, as discussed in Section 3 below, with Section 4 stressing the same broad message as emphasized by Zhang (2019). Section 5 harvests another low-hanging fruit of the *ddc* formulation, since it provides an immediate explanation of the celebrated double robustness. Section 6 then ventures into a much harder area of engineering a more representative sub-sample out of a large non-representative sample, a worthwhile trade-off because data quality is far more important than data quantity (Meng, 2018), as briefly reviewed below.

3. A unifying strategy based on data defect correlation

In the setup of Wu (2022), for each individual $i,$ we have a set of attributes $A_i = \{y_i, \mathbf{x}_i\},$ where y is the attribute of interest, and $\mathbf{x}$ is auxiliary, which is useful in two ways. First, reducing the sampling bias due to non-probability sampling becomes possible when the non-probability mechanism can be (fully) explained by $\mathbf{x}$. Second, by taking advantage of the relationships between y_i and $\mathbf{x}_i,$ we can improve the efficiency of our estimation. As a starting point, Wu (2022) assumes that we have two data sources available, which we denote via two recording indicators, R and $R^*.$ The main source of the data is a non-probability sample, where we observe both y_i and $\mathbf{x}_i$ for $i \in S \equiv \{i : R_i = 1\},$ but the recording indicator R_i is determined by a mechanism uncontrolled by any (known) design probability. A second source is (assumed to be) a probability sample, where we observe $\mathbf{x}_i$ only, for $i \in S^* \equiv \{i : R_i^* = 1\}.$ This second sample provides information to estimate population auxiliary information that is useful for estimating population quantities about $y,$ such as its mean. Hence this setup is closely related to the setup where $S \cup S^* = \mathcal{N};$ see Tan (2013).

Now for any function $m(\mathbf{x}),$ let $z_i = y - m(\mathbf{x}_i), i \in \mathcal{N}.$ Clearly we can estimate the population mean $\bar{y}_N = \mathrm{E}_I(y_I)$ via estimating $\bar{z} = \mathrm{E}_I(z_I)$ and $\bar{m} = \mathrm{E}_I\left[m(\mathbf{x}_I)\right].$ From the second sample, $\bar{m}$ can be estimated unbiasedly since it involves $\mathbf{x}$ only. We therefore can focus on estimating $\bar{z},$ while recognizing that a more principled approach is to set up a likelihood or Bayesian model to estimate all unknown quantities jointly (Pfeffermann, 2017). Applying identity (2.2) with $G = z$ then tells us that our central task is to choose the weight $\{W_i, i \in S\}$ and/or the m function to miniaturize the *ddc* $\rho_{\tilde{R},z}.$ For our current discussion, it is easier to explain everything via the covariance

$$c_{\tilde{R},z} \equiv \mathrm{Cov}_I(\tilde{R}_I, z_I) = \mathrm{Cov}_I(W_I R_I, y_I - m(\mathbf{x}_I)) = \frac{1}{N}\sum_{i=1}^{N} W_i R_i (z_i - \bar{z}) \tag{3.1}$$

instead of the correlation $\rho_{\tilde{R},z}$ because $\mathrm{Cov}_I(\tilde{R}_I, z_I)$ is a bi-linear function in R_I and $z_I.$ However, $\rho_{\tilde{R},z},$ being standardized, is more appealing theoretically and for modelling purposes; see Sections 6 and 7.

The expression in (3.1) tells us immediately how to make it zero in expectations operationally, and in what sense conceptually. For whatever probability we impose on R_i (to be specified in late sections), let $\pi_i = \Pr(R_i = 1 \mid \mathbf{A}),$ which we assume will depend on A_i only. Then the linearity of the covariance operator implies that the average covariance with respect to the randomness in R_i is given by

$$\mathrm{E}[c_{\tilde{R},z} \mid \mathbf{A}] = \mathrm{Cov}_I\left(W_I \pi_I, y_I - m(\mathbf{x}_I)\right), \tag{3.2}$$

where $\mathbf{A} = \{A_i, i \in \mathcal{N}\}$. Similarly, if one is willing to posit a joint model for $\{(R_i, y_i), i \in \mathcal{N}\}$ conditioning on $\mathbf{X}$ in the independence form $\Pi_{i=1}^{N} P\left(R_i, y_i \,\middle|\, \mathbf{x}_i\right)$, then

$$\mathrm{E}[c_{\tilde{R},z} \mid \mathbf{X}] = \mathrm{Cov}_I \left(W_I \pi_I, \mathrm{E}(y_I \mid \mathbf{x}_I) - m(\mathbf{x}_I)\right). \tag{3.3}$$

Very intuitively, one can ensure a zero covariance or correlation between two variables by making either of them a constant. The two choices then would lead to respectively the quasi-randomization approach by making $W_I \pi_I \propto 1$ and the super-population approach by making $\mathrm{E}[y_I \mid \mathbf{x}_I] - m(\mathbf{x}_I)$ a constant (e.g., zero). The fact that either one is sufficient to render zero covariance (under the joint model) yields the double robustness, because it does not matter which one. But clearly these are not the only methods to achieve a zero correlation/covariance or double robustness, an emphasis of Kang and Schafer (2007) in their attempt to demystify the doubly robust approach (Robins, Rotnitzky and Zhao, 1994; Robins, 2000; Scharfstein, Rotnitzky and Robins, 1999). See also Tan (2007, 2010) for discussions and comparisons of an array of estimators, including those corresponding to only the quasi-randomization approach or only the super-population approach, some of them are doubly robust.

Indeed, because formula (2.2) is an identity for the actual error, any asymptotically unbiased (linear) estimators of the population mean must imply its corresponding *ddc* is asymptotically unbiased for zero, and vice versa, with respect to the randomness in R or in $\{R, y\}$. However, it is possible for *ddc* to be asymptotically unbiased for zero, without assuming any model is correctly specified – see Section 5 for an example. (This “double-plus robustness” is different from the “multiple robustness” of Han and Wang (2013), which still needs to assume the validity of at least one of the posited multiple models.) These two observations suggest that any general sufficient and necessary strategy for ensuring asymptotically consistent/unbiased (linear) estimators for the population mean would be equivalent to miniaturizing *ddc*.

As an example of a unified insight that otherwise might not be as intuitive, expression (3.2) suggests that we should include our estimate of π_I as a part of the predictor in the regression model $m(\mathbf{x}_I)$, since that can help to reduce the correlation between $W_I \pi_I$ and $z_I = y_I - m(\mathbf{x}_I)$, especially when we use constant weights W_I. Using $\hat{\pi}_I$ as a predictor for y is generally hard to motivate purely from the regression perspective, especially when we assume y and R are independent given $\mathbf{x}$ (typically a necessary condition to proceed, as discussed in the next section). However, expression (3.2) tells us that for the purpose of estimating the mean of y, it is not absolutely necessary to fit the correct regression model $m(\mathbf{x})$. Rather, it is sufficient to ensure the “residual” z_I is as uncorrelated with $W_I \pi_I$ as I varies. However, it is critically important to recognize that it is not sufficient to ensure zero or small correlation only among the observed data, because $\mathrm{Cov}_I (W_I \pi_I, z_I \mid R_I = 1)$ tells us little about $\mathrm{Cov}_I (W_I \pi_I, z_I \mid R_I = 0)$. In the setting of Wu (2022), our ability to extrapolate from $R_I = 1$ to $R_I = 0$ depends on the availability of the (independent) auxiliary data indexed by $R_I^* = 1$, which allow us to observe some x_I’s for which $R_I = 0$.

The strategy of including propensity estimates as a predictor has been found beneficial in related literature. For example, Little and An (2004) included the logit of $\hat{\pi}$ in their imputation model, and reported the inclusion enhanced the robustness of the imputed mean to the misspecification of the imputation model. The method was further developed and enhanced by Zhang and Little (2009) and by Tan, Flannagan and Elliott (2019), who used the term "Robust-squared" to emphasize the enhanced robustness. In a more recent article on such a strategy for non-probability samples, Liu et al. (2021) emphasized the importance of including the estimated propensity $\hat{\pi}_i$ "as a predictor" in $m(x, \hat{\pi})$ (using notation in this article). Furthermore, in the literature of targeted maximum likelihood estimation (TMLE) for semi-parametric models for dealing with non-probability data (van der Laan and Rubin, 2006; Luque-Fernandez, Schomaker, Rachet and Schnitzer, 2018) (also see Scharfstein et al. (1999); Tan (2010)), the variables $R_I / \hat{\pi}_I$ and $(1-R_I)/(1-\hat{\pi}_I)$ are called *clever covariates* and are used in the regression models for y_I. The implementations and theories of TMLE, and the related Collaborative TMLE (van der Laan and Gruber, 2009, 2010), are mathematically more involved than those under finite-population settings as discussed below, but the insights gained from (3.2)-(3.3) can provide us with helpful intuitions on understanding the essence of such methods.

4. Quasi-randomization *or* super-population implementations

In a nutshell, the quasi-randomization approach focuses on making $W_I \pi_I$ a constant variable (induced by FPI I). When our sample is genuinely selected by a probabilistic scheme by design, then $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i),$ for $i \in \mathcal{N},$ is a design probability, free of $y_i,$ but it can depend on $\mathbf{x}_i$ for example when $\mathbf{x}_i$ includes a stratifying variable. When the design probability is unavailable, we first need to invoke a divine probability. This could be a natural one given by the finite population, such as the propensity $\pi_i = \Pr_I(R_I = 1 \mid A_I = A_i)$ induced by FPI, where $A_i = \{y_i, \mathbf{x}_i\},$ or an imagined super-population one such as the R_i's being generated independently from $\mathrm{Ber}(\pi_i),$ where $\pi_i = \Pr(R_i = 1 \mid A_i) > 0.$ This positivity assumption is necessary if the finite population is pre-specified, or its imposition defines the finite population that can be studied. (This is a practically rather relevant consideration, such as in election polling, where the finite population may not be always pre-specified even theoretically.) Since these divine probabilities are unknown and serve as our estimand, we need to assume some device probabilities, such as via a generalized linear model $\pi_i = g(y_i, \mathbf{x}_i)$ to proceed, even though we don't really believe in any particular choice of $g.$

For our current discussion, suppose our divine probability is given by the super-population Bernoulli model. Let $n_R = \sum_{i=1}^{N} R_i,$ and $\tilde{p}(\mathbf{A}) = \Pr(n_R > 0 \mid \mathbf{A}) = 1 - \Pi_{i \in N}(1-\pi_i),$ where $\mathbf{A} = \{A_i, i \in \mathcal{N}\}.$ Because the R_i here is controlled by a divine probability, the sample size n_R is no longer a design variable to be conditioned upon in our replication scheme; it is generally no longer an ancillary statistic. Nevertheless, we should condition on $n_R > 0,$ a universal requirement for constructing data-driven estimates for $\bar{G}.$ Fortunately this conditioning does not create mathematical complications to the simplicity granted by the independence among $\pi_i, i \in \mathcal{N}$ as functions of $A_i.$ This is because $\tilde{\pi}_i(\mathbf{A}) \equiv \Pr(R_i = 1 \mid \mathbf{A}, n_R > 0) = \pi_i / \tilde{p}(\mathbf{A}),$ but the normalizing constant $\tilde{p}(\mathbf{A})$ – which depends on the entire $\mathbf{A}$ – is not relevant for the developments in this article, such as assigning weights that are proportional to $\tilde{\pi}_i^{-1}(\mathbf{A})$.

Consequently, under this divine probability, which corresponds to (the true model for) the q-model setting in Wu (2022), we have for any chosen W_I, by (3.1)

$$\begin{aligned} \mathrm{E}\,(c_{\tilde{R},z} \,|\, \mathbf{A}, n_R > 0) &= \mathrm{Cov}_I\,(W_I \mathrm{E}[R_I \,|\, \mathbf{A}, n_R > 0], y_I - m(\mathbf{x}_I)) \\ &= \tilde{p}^{-1}(\mathbf{A})\,\mathrm{Cov}_I\,(W_I \pi_I, y_I - m(\mathbf{x}_I)), \end{aligned} \tag{4.1}$$

where E is with respect to the (unknown) divine probability over R_I (for fixed I). It follows then that, regardless of whether we want to ensure zero expectation in (3.2) or in (4.1), we will impose $W_I \pi_I \propto 1$, that is, $W_I \propto \pi_I^{-1}$, the well-known inverse probability weighting. Therefore, if our postulated model q permits us to reliably capture π_i in reality, then $c_{\tilde{R},z} = O_p\,(N^{-1/2})$ because it has mean zero (with respect to the divine probability), and it is a weighted average of N essentially independent Bernoulli variables, as seen in (3.1).

This is a randomization oriented approach because it treats the entire finite population attribute values $\mathbf{A}$ as fixed, and the hypothetical replications are generated only by repeated realizations of the recording indicator R_I. Of course, in general, the values of $\{\pi_i, i \in \mathcal{N}\}$ are unknown, and worse they are inestimable from a non-probability sample without further assumptions. To proceed, we pose assumptions such as missing at random, i.e., $\Pr(R_i = 1 \,|\, A_i) = \Pr(R_i = 1 \,|\, \mathbf{x}_i)$, and the requirement of an auxiliary sample so that we have some values of $\mathbf{x}_i$ with $R_i = 0$. We also have choices on how to estimate the inclusion propensity $\pi_i = \Pr(R_i = 1 \,|\, \mathbf{x}_i)$, parametrically or non-parametrically. These assumptions, requirements, and estimation methods are all essential for practical implementation, as carefully reviewed and discussed by Wu (2022); also see Tan (2010) for a detailed comparison of various estimation strategies. Nevertheless, the overarching idea of quasi-randomization methods is to choose W_I to free $\tilde{R}_I = W_I R_I$ from I in expectation over the posited hypothetical replications, to regain the freedom guaranteed by probability sampling.

Complementarily, the super-population approaches aim to miniaturize $c_{\tilde{R},z}$ via making the other variable in $c_{\tilde{R},z}$, that is, z_I free of I in expectation, but over a different hypothetical replication scheme. Here the idea is to choose an $m(\mathbf{x}_i)$ that is a good approximation to y_i such that the residual $z_i = y_i - m(\mathbf{x}_i)$ will be zero in expectation conditioning on $\mathbf{x}$. Typically, this is done by considering a joint model for $\{R_i, y_i\}$ given $\mathbf{x}_i$, and with a specific regression model $\xi\,(y \,|\, \mathbf{x})$, using the notation in Wu (2022). It is important to recognize that, although we only specify the regression model y_i given $\mathbf{x}_i$, we must include R_i in the replications in order to capture the possible dependence of R_i on the entire $A_i = \{y_i, \mathbf{x}_i\}$, which is the key concern for non-probability samples. Indeed, it is this joint specification that permits the adoption of the missing at random assumption to reduce $P(y_i \,|\, \mathbf{x}_i, R_i) = P(y_i \,|\, \mathbf{x}_i)$, which in turn permits us to focus on specifying a single regression model $\xi\,(y_i \,|\, \mathbf{x}_i)$ for both observed and unobserved individuals. Therefore, when we write E_ξ, we mean the expectation with respect to

$$P(R_i, y_i \mid \mathbf{x}_i) = P(R_i \mid \mathbf{x}_i) P(y_i \mid R_i, \mathbf{x}_i) = \pi_i^{R_i} (1-\pi_i)^{1-R_i} \xi(y_i \mid \mathbf{x}_i), \tag{4.2}$$

where $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i)$ is left unspecified, unlike with the quasi-randomization approach.

It follows then that, conditioning on $\mathbf{X} = \{\mathbf{x}_i, i \in \mathscr{N}\}$ and $n_R > 0,$ which does not alter $P(y \mid \mathbf{X})$ because y and R are independent given $\mathbf{X},$ we have

$$\mathrm{E}(c_{\tilde{R},z} \mid \mathbf{X}, n_R > 0) = [\tilde{p}(\mathbf{X})]^{-1} \mathrm{Cov}_I (W_I \pi_I, \mathrm{E}[y_I \mid \mathbf{x}_I] - m(\mathbf{x}_I)). \tag{4.3}$$

Clearly, (4.3) becomes zero when we choose $m(\mathbf{x}_I) = \mathrm{E}_\xi [y_I \mid \mathbf{x}_I]$ and that the ξ model is (first-order) correctly specified, that is, $\mathrm{E}_\xi [y_I \mid \mathbf{x}_I] = \mathrm{E}[y_I \mid \mathbf{x}_I].$ This summarizes the super-population approach, and it renders $c_{\tilde{R},z} = O_p (N^{-1/2})$ for similar reasons as given for the quasi-randomization framework.

5. Quasi-randomization *and* super-population implementations

Once a joint model for $\{R_i, y_i\}$ is set up, of course we can use it for estimating both π_i and the regression function $m(\mathbf{x}),$ each of which is made possible by the availability of the auxiliary probability sample, and the assumption of missing at random. But as shown before, correctly specifying and estimating one of them is sufficient for miniaturizing $c_{\tilde{R},z}.$ However, from (4.3), in order for the covariance/correlation to be zero, neither multiplicative correction to π_I via W_I nor the additive adjustment for $\mathrm{E}(y_I \mid \mathbf{x}_I)$ via $m(\mathbf{x}_I)$ need to be correct. All we need is that, after the correction or adjustment, what is left would be uncorrelated with each other. The aforementioned framework of Collaborative TMLE was built essentially on this insight (e.g., see Section 3.1 of van der Laan and Gruber, 2009), though the heavy mathematical treatments in its literature might have discouraged readers to seek such intuitive understanding.

To provide a simple illustration, consider a finite population that is an i.i.d. sample from a super-population model:

$$\mathrm{E}[y \mid x] = \sum_{k=0}^{3} \beta_k x^k, \quad x \sim N(0,1). \tag{5.1}$$

The non-probability sample is generated by a mechanism R such that $\Pr\big(R = 1 \mid y, x\big) = \pi\big(\,|x|\,\big),$ that is, it is determined by the magnitude of x only. Suppose we mis-specify the function form for π (e.g., the divine model may not be monotone in $|x|,$ but the device model such as the conventional logistic link is), as well the regression model by choosing $m(x) = b_0 + b_1 x + b_2 x^2.$ Since x^2 is uncorrelated with x or x^3 under $x \sim N(0,1),$ we know that our least-square estimator for b_2 would still be valid for β_2 even under the mis-specified regression model. This turns out to be sufficient to ensure the asymptotic unbiasedness (as $N \to \infty$) of the following “doubly robust” estimator for $\mu = \bar{y}_N,$ the finite-population mean,

$$\hat{\mu}_{+} = \frac{\sum_{i=1}^{N} R_i w\left(\left|x_i\right|\right)(y_i - \hat{m}(x_i))}{\sum_{i=1}^{N} R_i w\left(\left|x_i\right|\right)} + \frac{\sum_{i=1}^{N} R_i^{*} \hat{m}(x_i)}{\sum_{i=1}^{N} R_i^{*}}, \tag{5.2}$$

where R^{*} indicates the auxiliary sample (of $\mathbf{x}$ only). Or equivalently,

$$\hat{\mu}_{+} - \bar{y}_N = \frac{\mathrm{Cov}_I\left(R_I w\left(\left|x_I\right|\right), y_I - \hat{m}\left(x_I\right)\right)}{\mathrm{E}_I\left(R_I w\left(\left|x_I\right|\right)\right)} + \frac{\mathrm{Cov}_I\left(R_I^{*}, \hat{m}\left(x_I\right)\right)}{\mathrm{E}_I\left(R_I^{*}\right)}, \tag{5.3}$$

which makes it clearer that any bias in $\hat{\mu}_{+}$ is controlled by the covariance (or correlation) involving R, since the covariance involving R^{*} is already miniaturized by the assumption that the auxiliary sample is probabilistic (which, for simplicity, is assumed to be a simple random sample).

Here $w(x)$ is any weight function such that $\mathrm{E}_{\phi}\left[\left|x\right|^{3} w\left(\left|x\right|\right)\right] < \infty$, where the expectation is with respect to $x \sim N(0, 1)$, and $\hat{m}(x) = b_0 + b_1 x + \hat{\beta}_2 x^2$, with $\hat{\beta}_2$ being the least-square estimator for β_2 from the biased sample, and b_0 and b_1 can be chosen arbitrarily. Because the finite-population covariance/correlation between $\pi\left(\left|x_I\right|\right) w\left(\left|x_I\right|\right)$ and x_I^k is $O_p(N^{-1/2})$, for $k = 1$ and $k = 3$, the misfitted parts for π or m do not contribute to the *ddc* (asymptotically) since they are uncorrelated with each other under the super-population model, leading to further robustness going beyond “double robustness”. This of course does not mean that we can misfit a model arbitrarily and still obtain valid estimators, but it does imply that having at least one model being correct is a sufficient, but not necessary, condition for the validity of the doubly robust estimators.

It is also worth stressing that, in formatting the regression model, we do not necessarily need to invoke a device probability, e.g., a super-population regression model, because the FPI variable provides a finite-population regression via applying the least-squares method to regress y_i on $\mathbf{x}_i, i \in \mathcal{N}$. This regression fitting itself says little about whether the resulting regression line $y = \hat{m}(\mathbf{x})$ is a good fit to $(y_i, \mathbf{x}_i)$ or not. However, the example above indicates that, for the purpose of estimating the population average of y, the lack of fit may not matter that much, as long as the “residual” $z_I = y_I - \hat{m}(\mathbf{x}_I)$ has little correlation with $W_I \pi_I$, as two functions of the FPI variable I. Indeed, as discussed in Section 3, we can consider including $\hat{\pi}_I$ in the regression model $\hat{m}(\mathbf{x}_I, \hat{\pi}_I)$. How effective this strategy is in general is a topic of further research.

6. Counterbalancing sub-sampling

6.1 The devastating impact of data defect on effective sample size

A key finding, which has surprised many, from studying the data quality issue is how small the size of our “big data” is when we take into account the data defect. To prove this mathematically, we can equate the mean-squared error (MSE) of $\bar{G}_W$ in (2.1), with the MSE of a simple random sampling estimator of size n_{eff}. This yields (see Meng (2018) for derivation):

$$n_{\mathrm{eff}} \approx \frac{f_W}{1-f_W}\frac{1}{\mathrm{E}[\rho^2_{\tilde{R},G}]} \approx \frac{f_W}{1-f_W}\frac{1}{\rho^2_{\tilde{R},G}}, \tag{6.1}$$

where $f_W = n_W / N$ and the expectation E is with respect to the conditional distribution of $\tilde{R}$ given n_W. It is worthwhile to note that this (conditional) distribution can involve all three types of probability discussed in Section 1.2 because the variations in $\tilde{R}$ can come from multiple sources. For example, in typical opinion surveys, there will be (1) design probability in the sampling indicator, (2) divine probability in formulating the non-response mechanism, and (3) device probability for estimating the mechanism and the weights.

Expression (6.1) is the weighted version/extension of the expression given in Meng (2018) with equal weights, which reveals the devastating impact of a seemingly tiny *ddc*. Suppose our sample is 1% of the population, and it suffers from a half-percent *ddc*. Applying (6.1) (with equal weights) with $f_W = 0.01$ and $\rho_{\tilde{R},G} = 0.005$ yields $n_{\mathrm{eff}} \approx 404$ *regardless of the sample size* n_R. In the case of the 2020 US presidential election, 1% of the voting population is about 1.55 million people, and hence the loss of sample size due to a half percent *ddc* is about 1 - (404 / 1,550,000) > 99.97%. Such seemingly impossible losses have been reported in both election studies (Meng, 2018) and COVID vaccination studies (Bradley et al., 2021). A most devastating consequence of such losses is the "big data paradox": the larger the (apparent) data size, the surer we fool ourselves because our false confidence (in both technical and literal sense) goes up with the erroneous data size, while the actual coverage probability of the incorrectly constructed confidence intervals become vanishingly small (Meng, 2018; Msaouel, 2022).

A positive implication from this revelation, however, is that we can trade much data quantity for data quality, and still end up having statistically more accurate estimates. Of course, in order to reduce the bias, we will need some information about it. If we have reliable information on the value of *ddc*, we can directly adjust for the bias in estimating the population average corresponding to the *ddc*, for example by a Bayesian approach, similar to that taken by Isakov and Kuriwaki (2020) in their scenario analysis. Furthermore, if we have sufficient information to construct reliable weights, we can use the weights to adjust for selection biases as commonly done. Nevertheless, even in such cases, it may still be useful to create a representative miniature of the population out of a biased sample for general purposes, which for example can eliminate many practitioners' anxiety and potential mistakes for not knowing how to properly use the weights. Indeed, few really know how to deal with weights, because "Survey weighting is a mess" (Gelman, 2007).

However, creating a representative miniature out of a biased sample in general is a challenging task, especially because *ddc* can (and will) vary with the variable of interest. Nevertheless, just as weighting is popular tool despite it being far from perfect, let us explore representative miniaturization and see how far we can push the idea. The following example therefore is purely for brainstorming purposes, by looking into a common but challenging scenario, where we have reasonable information or understanding on the direction of the bias, that is, the sign of the *ddc*, but rather vague information about its magnitude. A good example is non-representativeness of election polls because voters tend to not want to disclose their preferences when they plan to vote for a socially unpopular candidate; we therefore know the direction of the bias, but not much about its degree other than some rough guesses (e.g., a range of 10 percentage points).

6.2 Creating a less biased sub-sample

The basic idea is to use such partial information about the selection bias to design a *biased* sub-sampling scheme to *counterbalance* the bias in the original sample, such that the resulting sub-samples have a *high likelihood* to be less biased than the original sample from our target population. That is, we create a sub-sampling indicator S_I, such that with high likelihood, the correlation between $S_I R_I$ and G_I is reduced, compared to the original $\rho_{R,G}$, to such a degree that it will compensate for the loss of sample size and hence reduce the MSE of our estimator (e.g., the sample average). We say with *high likelihood*, in its non-technical meaning, because without full information on the response/recording mechanism, we can never guarantee such a counterbalance sub-sampling (CBS) would always do better. However, with judicious execution, we can reduce the likelihood of making serious mistakes.

To illustrate, consider the case where y is binary. Let $\Delta = r_1 - r_0$, where r_y is the propensity of responding/reporting for individuals whose responses will take value y: $r_y = \mathrm{Pr}_I\,(R_I = 1 \mid y_I = y)$. If the sample is representative, then like $\rho_{R,G}$, Δ is miniaturized, meaning that it is on the order of $N^{-1/2}$. This is most clearly seen via the easily verifiable identity (see (4.1) of Meng, 2018)

$$\Delta = \frac{\mathrm{Cov}_I\,(y_I, R_I)}{p(1-p)} = \rho_{R,y}\sqrt{\frac{f_R\,(1-f_R)}{p(1-p)}}, \tag{6.2}$$

where $p = \mathrm{Pr}_I\,(y_I = 1)$ and $f_R = \mathrm{Pr}_I\,(R_I = 1)$, which is the original sampling rate. A key ingredient of CBS is to determine $s_y = P_I\,(S_I = 1 \mid y_I = y, R_I = 1)$ for $y = 0, 1$, that is, the sub-sampling probabilities of individuals who reported $y = 1$ and $y = 0$, respectively.

To determine the beneficial choices, let $f_S = \mathrm{Pr}_I\,(S_I = 1 \mid R_I = 1)$ be the sub-sampling rate, and $\Delta_S = s_1 r_1 - s_0 r_0$. Then by applying (2.2) (with equal weights) and (6.2) to both the sample average and the sub-sample average, we see that the sub-sample average has smaller (actual) error in magnitude if and only if

$$\left(\frac{\Delta_S}{f_S\,f_R}\right)^2 < \left(\frac{\Delta}{f_R}\right)^2 \Leftrightarrow f_S^2 > \left(\frac{\Delta_S}{\Delta}\right)^2. \tag{6.3}$$

Writing $r = r_1/r_0$ and $s = s_1/s_0$, the right-hand side of (6.3) becomes

$$\left[sp^* + (1-p^*)\right]^2 > \left(\frac{rs-1}{r-1}\right)^2, \tag{6.4}$$

where $p^* = \Pr_I(y_I = 1 | R_I = 1)$ is observed in the original sample, which should remind us that p^* may be rather different from the p we seek, because of the biased R-mechanism.

An immediate choice to satisfy (6.4) is to set $s = r^{-1}$, which of course typically is unrealistic because if we know the value of r, then the problem would be a lot simpler. To explore how much leeway we have in deviating from this ideal choice, let $\delta = r - 1$, we can then show that (6.4) is equivalent to

$$(s-1)\left\{[1+(1+p^*)\,\delta]\,(s-1)+2\delta\right\} < 0. \tag{6.5}$$

This tells precisely the permissible choices of s without over-correcting (in the magnitude of the resulting bias):

(i) When $r > 1$, i.e., $\delta > 0$, we can take any s such that

$$\frac{[1-(1-p^*)\,\delta]_+}{1+(1+p^*)\,\delta} \le s < 1; \tag{6.6}$$

(ii) When $r < 1$, i.e., $\delta < 0$, we can take any s such that

$$1 < s \le \frac{1-(1-p^*)\,\delta}{[1+(1+p^*)\,\delta]_+}. \tag{6.7}$$

This pair of results confirms a number of our intuitions, but also offers some qualifications that are not so obvious. Since we sub-sample to compensate for the bias in the original sample, s and r must stay on the opposite side of 1, i.e., $(s-1)(r-1) = (s-1)\,\delta < 0$, as seen in (6.6)-(6.7). To prevent over corrections, some limits are needed, but it is also possible that the initial bias is so bad that no sub-sampling scheme can make things worse, which is reflected by the positivizing function $[x]_+$ in the two expressions above. However, the expressions for the limits as well as for the thresholds to activate the positivizing functions are not so obvious. Nor is it obvious that these expressions depend on the unknown p indirectly via the observed p^*, and hence only prior knowledge of r is required for implementing or assessing CBS.

This observation suggests that it is possible to implement a beneficial CBS when we can borrow information from other surveys (or studies) where the response/recording behaviors are of similar nature. For example, we may learn that a previous similar survey had $r =$ 1.5 (e.g., those with $y = 1$ had 6% of chance to be recorded, and those with $y = 0$ had only 4% chance). Taking into account the uncertainty in the similarity between the two surveys, we might feel comfortable to place (1.2, 1.8) as the plausible range for r in the current study. Suppose we observe $p^* =$ 0.6, this means that the maximum – over the range $r \in$ (1.2, 1.8) – of the lower bound on the permissible s as given in (6.6) is

$$\frac{[1-(1-0.6)\,(r-1)]_+}{1+1.6(r-1)} = \frac{[1.4-0.4r]_+}{1.6r-0.6} \le \frac{1.4-0.4\times 1.2}{1.6\times 1.2-0.6} = 0.7. \tag{6.8}$$

Therefore, as long as we choose $s \in [0.7, 1)$, we are unlikely to over-correct. The price we pay for this robustness is that the resulting sub-sample is not as good quality as it can be, for example, when the underlying r for the current study is indeed 1.5 (in expectation). Choosing any $s \in [0.7, 1)$ will not provide the full correction as provided by $s = 1/r = 0.67$, that is, the sub-sample average will still have a positive bias but with a smaller MSE compared to the original sample average. Of course both the feasibility and effectiveness of such CBS need to be carefully investigated before it can be recommended for general consumption, especially going beyond binary y. The literature on inverse sampling (Hinkins, Oh and Scheuren, 1997; Rao, Scott and Benhin, 2003) is of great relevance for such investigations, because it also aims to produce simple random samples via subsampling, albeit with a different motivation (to turn complex surveys into simple ones for ease of analysis).

7. Probability sampling as aspiration, not prescription

As it should be clear from the definition of *ddc*, it is not directly estimable from the biased sample alone. One therefore naturally would (and should) question how useful *ddc* is or could be. The answer turns out to be an increasingly long one thanks to *ddc* being model-free and hence a versatile data quality metric for both probability samples and non-probability samples. Its usefulness for generating theoretical insights is demonstrated by its role in quantifying the data quality-quantify trade-off via effective sample size as seen in (6.1), in understanding simulation errors in quasi-Monte Carlo as explored in Hickernell (2016), and in anticipating the “double-plus robustness” phenomenon as presented in Section 5. Its methodological usages are illustrated by the scenario analyses for the 2020 US Presidential election (Isakov and Kuriwaki, 2020) and for the COVID-19 vaccination assessments (Bradley et al., 2021). Its practical implications can be found in epidemiological studies (Dempsey, 2020), particle physics (Courtoy, Houston, Nadolsky, Xie, Yan and Yuan, 2022), and political polling (Bailey, 2023).

Not surprisingly, these practical applications found the notion of *ddc* and the underlying error decomposition (2.2) helpful because of the non-probability samples they need to deal with, either due to distortions to the probability samples such as by a biased non-response mechanism or due to selection biases in the first place such as selective COVID-19 testing. Professor Wu’s overview, and the many references cited there and in this discussion, should make it clear that non-probability samples are *almost surely* everywhere. I am invoking this strong probabilistic phrase not merely for its humorous value. When we consider the unaccountably many possible values for the mean of *ddc*, the probability – however we construct it to capture the wild west of data collection processes out there – that it will land precisely on zero must be zero. This zero mean is a necessary condition for the sample to be a probability sample, because a probability sample implies that *ddc* must be of the order of $N^{-1/2}$ order (Meng, 2018), which is impossible when its mean is non-zero (asymptotically). This observation suggests that we should move away from our tradition of treating probability sampling as a centerpiece and then try to model the much larger world of non-probability samples as “deviations” from it. Instead, we should start with studying samples with general collection mechanisms using tools or concepts such as *ddc*, and then treat (design) probability samples as the very special, ideal case – always an aspiration, but never the only prescription for action.

Acknowledgements

I am grateful to Editor Jean-François Beaumont for inviting me to discuss Changbao Wu's timely and thought-provoking overview. I thank James Bailie, Radu Craiu, Adel Daoud, Andrew Gelman, Stas Kolenikov, Rod Little, Cory McCartan, Kelly McConville, James Robins, Zhiqiang Tan, and Li-Chun Zhang for moral endorsement and for constructive criticisms. I also thank NSF for partial financial support, and Steve Finch for careful proofreading.

References

Bailey, M.A. (2023). *Polling at a Crossroads – Rethinking Modern Survey Research*. Cambridge University Press.

Beaumont, J.-F., and Rao, J.N.K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *Survey Statistician*, 83, 11-22.

Blei, D.M., Kucukelbir, A. and McAuliffe, J.D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 859-877.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, Z.-L. and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890), 695-700.

Buelens, B., Burger, J. and van den Brakel, J.A. (2018). Comparing inference methods for nonprobability samples. *International Statistical Review*, 86(2), 322-343.

Courtoy, A., Houston, J., Nadolsky, P., Xie, K., Yan, M. and Yuan, C.-P. (2022). Parton distributions need representative sampling. *arXiv preprint arXiv:2205.10444.*

Craiu, R.V., Gong, R. and Meng, X.-L. (2022). Six statistical senses. *arXiv preprint arXiv:2204.05313.*

David Peat, F. (2002). *From Certainty to Uncertainty: The Story of Science and Ideas in the Twentieth Century*. Joseph Henry Press.

Dempsey, W. (2020). The hypothesis of testing: Paradoxes arising out of reported coronavirus case-counts. *arXiv preprint arXiv:2005.10425*.

Dwork, C. (2008). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, Springer, 1-19.

Elliott, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2), 153-164.

Gong, R. (2022). Transparent privacy is principled privacy. *Harvard Data Science Review*, (Special Issue 2), June 24, 2022. https://hdsr.mitpress.mit.edu/pub/ld4smnnf.

Gong, R., Groshen, E.L. and Vadhan, S. (2022). Harnessing the known unknowns: Differential privacy and the 2020 Census. *Harvard Data Science Review*, (Special Issue 2), June 24 2022. https://hdsr.mitpress.mit.edu/pub/fgyf5cne.

Han, P., and Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, 100(2), 417-430.

Hartley, H.O., and Ross, A. (1954). Unbiased ratio estimators. *Nature*, 174(4423), 270-271.

Hickernell, F.J. (2016). The trio identity for Quasi-Monte Carlo error. In International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, Springer, 3-27.

Hinkins, S., Oh, H.L. and Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 1, 11-21. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3101-eng.pdf.

Isakov, M., and Kuriwaki, S. (2020). Towards principled unskewing: Viewing 2020 election polls through a corrective Lens from 2016. *Harvard Data Science Review*, 2(4), Nov. 3, 2020. https://hdsr.mitpress.mit.edu/pub/cnxbwum6.

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. *Statistical Science*, 22(4), 523-539.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Li, X., and Meng, X.-L. (2021). A multi-resolution theory for approximating infinite-*p*-zero-*n*: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association*, 116(533), 353-367.

Little, R., and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14(3), 949-968.

Liu, Y., Gelman, A. and Chen, Q. (2021). Inference from non-random samples using Bayesian machine learning. *arXiv preprint arXiv:2104.05192*.

Lo, A.W. (2017). Adaptive markets. In *Adaptive Markets*. Princeton University Press.

Lohr, S., and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101(475), 1019-1030.

Lohr, S.L. (2021). *Sampling: Design and Analysis*. Chapman and Hall/CRC.

Luque-Fernandez, M.A., Schomaker, M., Rachet, B. and Schnitzer, M.E. (2018). Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, 37(16), 2530-2546.

Meng, X.-L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). In *Past, Present, and Future of Statistical Science*, (Eds., Lin et al.), CRC Press.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i) Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2), 685-726.

Meng, X.-L. (2021). Enhancing (publications on) data quality: Deeper data minding and fuller data confession. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(4), 1161-1175.

Msaouel, P. (2022). The big data paradox in clinical practice. *Cancer Investigation*, 1-27.

Pfeffermann, D. (2017). Bayes-based non-bayesian inference on finite populations from non-representative samples: A unified approach. *Calcutta Statistical Association Bulletin*, 69(1), 35-63.

Rao, J.N.K., Scott, A.J. and Benhin, E. (2003). Undoing complex survey data structures: Some theory and applications of inverse sampling. *Survey Methodology*, 29, 2, 107-128. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003002/article/6787-eng.pdf.

Robins, J.M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, Indianapolis, IN, 1999, 6-10.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846-866.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussions). *Journal of the American Statistical Association*, 94(448), 1096-1146.

Slavkovic, A., and Seeman, J. (2022). Statistical data privacy: A song of privacy and utility. *arXiv preprint arXiv:2205.03336*.

Tan, Y.V., Flannagan, C.A.C. and Elliott, M.R. (2019). “Robust-Squared” imputation models using Bart. *Journal of Survey Statistics and Methodology*, 7(4), 465-497.

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science*, 22(4), 560-568.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3), 661-682.

Tan, Z. (2013). Simple design-efficient calibration estimators for rejective and high-entropy sampling. *Biometrika*, 100(2), 399-415.

Van Buuren, S., and Oudshoorn, K. (1999). *Flexible Multivariate Imputation by MICE*. Leiden: TNO.

van der Laan, M.J., and Gruber, S. (2009). Collaborative double robust targeted penalized maximum likelihood estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, 246.

van der Laan, M.J., and Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1).

van der Laan, M.J., and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), 980-991.

Wu, C. (2022). Statistical inference with non-probability survey samples (with discussions). *Survey Methodology*, 48, 2, 283-311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf.

Wu, C., and Thompson, M.E. (2020). *Sampling Theory and Practice*. Springer.

Yang, S., Kim, J.K. and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2), 445-465.

Zhang, G., and Little, R. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65(3), 911-918.

Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields*, 3(2), 103-113.