The Annals of Applied Statistics
2018, Vol. 12, No. 2, 685–726
https://doi.org/10.1214/18-AOAS1161SF

STATISTICAL PARADISES AND PARADOXES IN BIG DATA (I): LAW OF LARGE POPULATIONS, BIG DATA PARADOX, AND THE 2016 US PRESIDENTIAL ELECTION[1]

BY XIAO-LI MENG

Harvard University

Statisticians are increasingly posed with thought-provoking and even paradoxical questions, challenging our qualifications for entering the statistical paradises created by Big Data. By developing measures for data quality, this article suggests a framework to address such a question: "Which one should I trust more: a 1% survey with 60% response rate or a self-reported administrative dataset covering 80% of the population?" A 5-element Euler-formula-like identity shows that for any dataset of size n, probabilistic or not, the difference between the sample average $\overline{X}_n$ and the population average $\overline{X}_N$ is the product of three terms: (1) a *data quality* measure, $\rho_{R,X}$, the correlation between X_j and the response/recording indicator R_j; (2) a *data quantity* measure, $\sqrt{(N-n)/n}$, where N is the population size; and (3) a *problem difficulty* measure, σ_X, the standard deviation of X. This decomposition provides multiple insights: (I) Probabilistic sampling ensures high data quality by controlling $\rho_{R,X}$ at the level of $N^{-1/2}$; (II) When we lose this control, the impact of N is no longer canceled by $\rho_{R,X}$, leading to a *Law of Large Populations* (LLP), that is, our estimation error, relative to the benchmarking rate $1/\sqrt{n}$, increases with $\sqrt{N}$; and (III) the "bigness" of such Big Data (for population inferences) should be measured by the *relative size* $f = n/N$, not the *absolute size* n; (IV) When combining data sources for population inferences, those relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.

Estimates obtained from the Cooperative Congressional Election Study (CCES) of the 2016 US presidential election suggest a $\rho_{R,X} \approx -0.005$ for self-reporting to vote for Donald Trump. Because of LLP, this seemingly minuscule data defect correlation implies that the simple sample proportion of the self-reported voting preference for Trump from 1% of the US eligible voters, that is, $n \approx 2{,}300{,}000$, has the same mean squared error as the corresponding sample proportion from a genuine simple random sample of size $n \approx 400$, a 99.98% reduction of sample size (and hence our confidence). The

Received December 2017; revised April 2018.

[1]I thank the Editors Tilmann Gneiting and Karen Kafadar for inviting this article for a special issue of *Annals of Applied Statistics* in memory of Stephen Fienberg, who taught me—among other things—how to prepare an impactful visiting committee report when I served on a visiting committee that he chaired to the Department of Statistics at the University of Virginia. (More acknowledgements appear at the end of the article. Feedback, regardless of their signs, are highly appreciated: E-mail: meng@stat.harvard.edu.)

Key words and phrases. Bias-variance tradeoff, data defect correlation, data defect index (d.d.i.), data confidentiality and privacy, data quality-quantity tradeoff, Euler identity, Monte Carlo and Quasi Monte Carlo (MCQMC), non-response bias.

CCES data demonstrate LLP vividly: on average, the larger the state's voter populations, the further away the actual Trump vote shares from the usual 95% confidence intervals based on the sample proportions. This should remind us that, without taking data quality into account, population inferences with Big Data are subject to a *Big Data Paradox*: the more the data, the surer we fool ourselves.

1. Prologue: Paradise gained or lost? Big Data, however the term is defined or (un)appreciated, has posed a paradoxical situation for Statistics and statisticians, in both external perceptions and internal reflections. Almost since the dawn of statistics, the dominating mathematical tool for justifying statistical methods has been *large-sample asymptotics*. Neither the Law of Large Numbers nor the Central Limit Theorem, two pillars of the vast statistical palace, could be established without such asymptotics. Surely then we statisticians must be thrilled by the explosive growth of data size, justifying all the large-sample asymptotic results out there. A statistical paradise would seem to have arrived.

The reality appears to be the opposite. The size of our data greatly exceeds the *volume* that can be comfortably handled by our laptops or software, and the *variety* of the data challenges the most sophisticated models or tools at our disposal. Many problems demand the type of *velocity* that would make our head spin for both data processing and analysis. But the worst of all, it appears that the more we lament how our nutritious recipes are increasingly being ignored, the more fast food is being produced, consumed and even celebrated as the cuisine of a coming age. Indeed, some of our most seasoned chefs are working tirelessly to preserve our time-honored culinary skills, while others are preparing themselves for the game of speed cooking. Yet others need a daily nightcap to ease the nightmare of the forever lost statistical paradise, even before it actually arrives.

Am I too pessimistic or cynical? I'll let you be the judge, as you know best to which group you belong. As for my group membership, I share the concern of paradise lost if too many of us are capable of only reminiscing about our (not too) good old days. But I see a paradise, or even paradises, gained if there is a sufficient number of us who can engage in what we have advertised to be the hallmark of our discipline, that is, principled thinking and methodology development for dealing with uncertainty. Fast food will always exist because of the demand—how many of us have *repeatedly* had those quick bites that our doctors have *repeatedly* told us stay away from? But this is the very reason that we need more people to work on understanding and warning about the ingredients that make fast food (methods) harmful; to study how to reduce the harm without unduly affecting their appeal; and to supply healthier and tastier meals (more principled and efficient methods) that are affordable (applicable) by the general public (users).

This is how I see paradises arising. Big Data have given us many fascinating and challenging research problems, for which we statisticians have multiple—yet unique—roles to play. To solve them well, typically a team is needed, consisting

of computer scientists, domain experts, (applied) mathematicians, statisticians, etc. Our century-long contemplation of principled assessments of uncertainty should, at the very least, help the team to greatly reduce unnecessary trials and errors by avoiding statistical fallacies, unrealistic expectations, misguided intuitions, and misleading evaluations. Re-inventing the wheel is a well-known phenomenon in almost any field and it is a common source of unhappiness in academia. But from a practical and societal point of view, the real damage occurs when the re-invented wheels are inferior, increasing the frequency of serious or even fatal accidents. Quality control is thus an important role for statisticians to carry out, as well as a force for innovation because real advances occur more from the desire to improve quality than quantity.

Indeed, the current project started when I was asked to help with statistical quality control by an agency. Among the first questions was "Which one should we trust more, a 5% survey sample or an 80% administrative dataset?", which led to the development of the *data defect index*, a main subject of this paper. Hence this paper focuses on population inferences from Big Data. The harder problem of individualized predictions with Big Data will be treated in the sequel, Meng (2018). For population inferences, a key "policy proposal" of the current paper is to shift from our traditional focus on assessing probabilistic uncertainty (e.g., in a sample mean) in the familiar form of

$$\text{Standard Error} \propto \frac{\sigma}{\sqrt{n}}$$

to the practice of ascertaining systematic error in non-probabilistic Big Data captured by

$$\text{Relative Bias} \propto \rho\sqrt{N}.$$

Here "Relative Bias" is the bias in the sample mean relative to a benchmarking standard error, σ and n are standard deviation and sample size, and N is the long forgotten *population size*. The unfamiliar term ρ is a *data defect correlation*, defined in this paper. We demonstrate via theoretical and empirical evidence that this shift is necessary if we want our error assessments—and our roles as experts on uncertainty quantifications—to remain relevant for Big-Data population inferences.

Specifically, Section 2 introduces a fundamental identity that quantifies the tradeoff between data quantity and data quality for using sample averages to estimate population averages, and correspondingly the concept of *data defect index*. Section 3 uses insights generated by the identity to reveal troubling phenomena arising from low-quality Big Data, especially a *Law of Large Populations* and a *Big Data Paradox*. Section 4 then applies these concepts and results to binary outcomes and to the 2016 US presidential election, which reveals a plausible explanation for our collective pre-election over-confidence (and hence post-election surprise). Section 5 makes a pedagogical link to the well-known Euler identity in mathematics, and discusses the use of the fundamental identity for Monte Carlo and Quasi Monte Carlo integrations, as well as for improving data confidentiality.

2. A fundamental identity for data quality-quantity tradeoff.

2.1. *Motivating questions*. "Which one should I trust more: a 1% survey with 60% response rate or a non-probabilistic dataset covering 80% of the population?" Such a question was posed, for example, by Professor Thomas Louis in his presentation of Keiding and Louis (2016) at the Joint Statistical Meetings (JSM) 2015 in Seattle. Raised prior to the arrival of the era of Big Data, this question would likely be treated as an academic curiosity—how often can we get a hold of 80% of a population? Isn't the whole idea of survey sampling to learn about a population without having to record a large chunk of it?

Indeed, learning reliably about a population via probabilistic sampling a soupçon of it was a revolutionary idea at the turn of the 19th century, an idea that took essentially half a century to be (almost) universally accepted; see Bethlehem (2009) and Fuller (2011), the latter of which is also a rich source of theory and methods in sampling surveys. A good way to explain this seemingly magical power is to analogize it to the tasting of soup—as long as the soup is stirred sufficiently uniformly, a spoonful is all it takes to ascertain the flavor of the soup regardless of the size of its container. A tiny high quality sample can provide as much information as many large ones with low quality, and here the quality is measured by the *representativeness*, achieved via "uniform stirring." For most human (and other) populations, "uniform stirring" is not feasible, but probabilistic sampling does essentially the same trick.

Therefore, the question raised above is about the tradeoff between data quantity and quality. This tradeoff is even clearer in a question raised in another presentation six years earlier: "*Integrated and Synthetic Data: Part of Future Statistics*?" by Dr. Jeremy Wu, then the Director of the LED (Local Employment Dynamics) project at the US Census Bureau. After reviewing the long history of surveys and how the study of statistics gained its vitality by showing "A 5% random sample is 'better' than 5% non-random sample in measurable terms", Dr. Wu asked, "Is an 80% non-random sample 'better' than a 5% random sample in measurable terms? 90%? 95%? 99%?"

The qualitative answer clearly is "it depends", on how non-random the larger sample is. We would imagine that a small departure from being random should not overwhelm the large gain in sample size. But how small must it be? And indeed how to quantify "better" or being "non-random"? The question raised by Professor Louis is even harder, because the quality of the probabilistic survey itself has been degraded by the non-response mechanism, typically a non-probabilistic process in itself, creating the well-known problem of a non-ignorable missing-data mechanism [Heitjan and Rubin (1990), Rubin (1976)]. Therefore a key question is *how to compare two datasets with different quantities and different qualities*?

Such questions become increasingly relevant as we venture deeper into the Big Data era, a signature of which is the availability of many datasets covering large percentages of their respective populations, yet they were never intended to be

probabilistic samples. For example, the LED project used unemployment insurance wage records, which cover more than 90% of the US workforce, and the records were kept because of law (but not the law of large numbers), and it is known to exclude all federal employers. It clearly would be foolish to ignore such big datasets because they are not probabilistic or representative. But in order to use them, we minimally need to know how much they can help or whether they can actually do more harm than help. The following development was built upon an earlier idea in Meng (2014), where an approximate identity was obtained because of the use of the propensity instead of the actual data recording indicator, as defined below.

2.2. *An identity linking data quantity, data quality and problem difficulty.* Let us start by considering a *finite population*, as in virtually all real-life problems, with individuals (not necessarily human) indexed by $j = 1, \ldots, N$. Suppose the individual attributes of interests are coded into a (multi-dimensional) variable X. As is well known, many population quantities of interest can be expressed as the population average of $\{G_j \equiv G(X_j), j = 1, \ldots, N\}$, denoted by $\overline{G}_N$, by choosing an appropriate function G, such as polynomial functions for moments and indicator functions for distributions or quantiles; for simplicity of notation, we will assume G maps X to the real line. Therefore, when we have a sample, say $\{X_j, j \in I_n\}$, where I_n is a size n subset of $\{1, \ldots, N\}$, the most routinely adopted estimator of $\overline{G}_N$ is the sample average [for good reasons as it is often both design consistent and model-based consistent; see Firth and Bennett (1998)],

$$\overline{G}_n = \frac{1}{n} \sum_{j \in I_n} G_j = \frac{\sum_{j=1}^N R_j G_j}{\sum_{j=1}^N R_j}, \tag{2.1}$$

where $R_j = 1$ for $j \in I_n$ and $R_j = 0$ otherwise. Here the letter "R", which leads to the *R-mechanism*, is used to remind ourselves of many possible ways that a sample arrived at our desk or disk, most of which are not of a probabilistic sampling nature. For *Random* sampling, $\boldsymbol{R} \equiv \{R_1, \ldots, R_N\}$ has a well-specified joint distribution, conditioning on the sample size $\sum_{j=1}^N R_j = n$. This is the case when we conduct probabilistic sampling and we are able to record all the intended data, typically unachievable in practice, other than with Monte Carlo simulations (see Section 5.3).

For many Big Data out there, however, they are either self-*Reported* or administratively *Recorded*, with no consideration for probabilistic sampling whatsoever. Even in cases where the data collector started with a probabilistic sampling design, as in many social science studies or governmental data projects, in the end we have only observations from those who choose to *Respond*, a process which again is outside of the probabilistic sampling framework. These "R-mechanisms" therefore are crucial in determining the accuracy of $\overline{G}_n$ as an estimator of $\overline{G}_N$; for

simplicity, hereafter the phrase "recorded" or "recording" will be used to represent all such R-mechanisms.

It is thus entirely possible that nothing in (2.1) is probabilistic. The X_j's and hence G_j's are fixed, as usual with a finite-population framework [see for example Royall (1968)]. The R_j's can be fixed as well, that is, no matter how often we repeat the process (as a thought experiment), each individual will either always choose to report or never report. This, however, does not render our beloved probabilistic approach obsolete. Far from it, a simple probabilistic argument provides a deep insight into how to quantify the recording mechanism, and how it affects the accuracy of $\overline{G}_n$.

The key here is to express the actual error $\overline{G}_n - \overline{G}_N$ in statistical terms that can generate insights. The standard tool of expressing a sample average as an expectation with respect to an empirical distribution comes in handy for this purpose. Specifically, for any set of numbers $\{A_1, \ldots, A_N\}$, we can view it as the support of a random variable A_J induced by the random index J defined on $\{1, \ldots, N\}$. When J is uniformly distributed, $\mathrm{E}_J(A_J) = \sum_{j=1}^N A_j/N \equiv \overline{A}_N$, the usual average. Consequently, the difference between $\overline{G}_n$ and $\overline{G}_N$ can be written as

$$\begin{aligned}\overline{G}_n - \overline{G}_N = \frac{\mathrm{E}_J(R_J G_J)}{\mathrm{E}_J(R_J)} - \mathrm{E}_J(G_J) &= \frac{\mathrm{E}_J(R_J G_J) - \mathrm{E}_J(R_J)\mathrm{E}_J(G_J)}{\mathrm{E}_J(R_J)} \\ &= \frac{\mathrm{Cov}_J(R_J, G_J)}{\mathrm{E}_J(R_J)},\end{aligned} \tag{2.2}$$

where E_J and Cov_J are all taken with respect to the uniform distribution on $J \in \{1, \ldots, N\}$. This is a trivial variation of the key identity for bounding the bias of ratio estimators [see Hartley and Ross (1954), Meng (1993)]. Yet it holds critical insights we need in order to quantify our estimation error with both probabilistic and non-probabilistic samples, that is, with any R-mechanism.

To see this, we first let $\rho_{R,G} = \mathrm{Corr}_J(R_J, G_J)$ be the (population) correlation between R_J and G_J, $f = \mathrm{E}_J(R_J) = n/N$ be the sampling rate, and σ_G be the standard deviation of G_J, all defined according to the uniform distribution of J. Then, using the fact that the variance of the binary R_J is $\mathrm{V}_J(R_J) = f(1-f)$, we have from (2.2) that

$$\overline{G}_n - \overline{G}_N = \underbrace{\rho_{R,G}}_{\text{Data Quality}} \times \underbrace{\sqrt{\frac{1-f}{f}}}_{\text{Data Quantity}} \times \underbrace{\sigma_G}_{\text{Problem Difficulty}}. \tag{2.3}$$

This identity tells us that there are three—and only three—factors that determine our estimation error. The obvious factor of *data quantity* is captured by $(1-f)/f$ in the second term on the right-hand side of (2.3), which renders precisely zero error when we record (accurately) the entire population ($f = 1$) and infinite error when we record no data ($f = 0$). Another obvious factor is the *problem difficulty*

captured by σ_G or equivalently by σ_G^2. If G_J is a constant (i.e., $\sigma_G^2 = 0$), then it is the easiest estimation problem, because $n = 1$ is sufficient to render zero error. The more variation among G_j's, that is, the larger the σ_G, the more difficult to estimate $\overline{G}_N$ accurately.

As we shall demonstrate via theoretical and empirical evidence throughout this paper, the most critical—yet most challenging to assess—among the three is *data quality*. Identity (2.3) establishes that for sample averages, the data quality is captured by the *data defect correlation* $\rho_{R,G}$ because it precisely measures both the sign and degree of selection bias caused by the R-mechanism. Intuitively, in the absence of any selection bias, such as under a genuine probabilistic sampling, the chance that a particular value of G is recorded/reported or not should not depend on the value itself. Consequently, $\rho_{R,G}$ should be zero on average (over the potential randomness in $\boldsymbol{R}$). On the other hand, if larger values of G have higher/lower chances to be recorded, then $\overline{G}_n$ overestimates/underestimates $\overline{G}_N$. Such tendency is indicated by the sign of $\rho_{R,G}$, and the degree of the bias is captured by the magnitude of $\rho_{R,G}$ (for given data quantity and problem difficulty).

It is important to emphasize that the identity (2.3) is exact and truly free of any mathematical or probabilistic assumption because the right-hand side of (2.3) is merely an algebraic re-expression of its left-hand side. Statistically, (2.3) is applicable whenever the recorded values of G can be trusted; for example, if a response is to vote for Clinton, it means that the respondent is sufficiently inclined to vote for Clinton at the time of response, not anything else. Otherwise we will be dealing with a much harder problem of *response bias*, which would require strong substantive knowledge and model assumptions [see, e.g., Liu et al. (2013)]. See Shirani-Mehr et al. (2018) for a discussion of other types of response bias that contribute to the so-called *Total Error* of survey estimates.

Under the assumption of no such response bias, (2.3) allows us to express the mean-squared error (MSE) of $\overline{G}_n$, $\mathrm{MSE}_{\boldsymbol{R}}(\overline{G}_n) = \mathrm{E}_{\boldsymbol{R}}[\overline{G}_n - \overline{G}_N]^2$, over any R-mechanism,[2] as

$$\mathrm{MSE}_{\boldsymbol{R}}(\overline{G}_n) = \mathrm{E}_{\boldsymbol{R}}\big[\rho_{R,G}^2\big] \times \left(\frac{1-f}{f}\right) \times \sigma_G^2 \equiv D_I \times D_O \times D_U, \tag{2.4}$$

where $\mathrm{E}_{\boldsymbol{R}}$ denotes the expectation with respect to any chosen distribution of $\boldsymbol{R}$ but conditioning on the sample size $\sum_{j=1}^{N} R_j = n$, as is typical with finite sample calculations. Here the notation of Ds—with subscripts "I O U" for easy memorization—is adopted both for simplicity and for potential generalizability going beyond estimating population means. For notation simplicity, we have suppressed, but do not forget, the dependence of D_I and D_U on G. Identity (2.4) reinforces the three ways of reducing the MSE:

[2]This includes the trivial case where $\boldsymbol{R}$ is deterministic, by using a singleton-mass probability distribution.

(I) *Increase the data quality* by reducing $D_I = \mathrm{E}_{\boldsymbol{R}}[\rho_{R,G}^2]$, a data *Defect Index*[3]—this is the aim of all probabilistic sampling designs, as we shall discuss in Section 3.1;

(II) *Increase the data quantity* by reducing the *Dropout Odds*: $D_O = (1-f)/f$—Big Data promises this, but unfortunately it is far less effective than reducing D_I—see Section 3.2;

(III) *Reduce the difficulty* of the estimation problem by reducing the *Degree of Uncertainty* $D_U = \sigma_G^2$—this is typically actionable only with additional information; see Section 5.1.

It is of critical importance to recognize explicitly that the concept of *data quality* must be a relative one, and more precisely it should be termed as *data quality for a particular study*. This is because any meaningful quantification of data quality, such as the data defect index (d.d.i.) D_I we just defined, must depend on (1) the purposes of the analysis—a dataset can be of very high quality for one purpose but useless for another (e.g., the choice of the G function[4]); (2) the method of analysis (e.g., the choice of sample average instead of sample median); and (3) the actual data the analyst includes in the analysis, which is an integrated part of the R-mechanism. As a demonstration, Section 3.4 will provide an identity that generalizes (2.3)–(2.4) to weighted estimators, which illustrates how data quality is affected by the weights.

We also emphasize that assessing the d.d.i. D_I is relevant even if we cannot determine whether the defects are mostly due to data collection or due to data analysis. This is because either way it can help inform future studies where a similar integrated process of data collection and analysis is repeated. Section 4.2 will illustrate this point in the context of the 2016 US general election, with the aim of gaining a 2020 vision for the next US presidential election. In the context of nonresponse or more generally missing data, d.d.i. can also be used as a measure of the degree of "nonignorability", in the same spirit of *index of sensitivity to nonignorability* (ISNI) proposed in Troxel, Ma and Heitjan (2004). The main difference is that D_I is defined with respect to sample mean and hence it requires no likelihood specification, whereas ISNI aims to measure the rate of change of the parametric likelihood inference as one moves away from the assumption of ignorability.

[3]In Meng (2014), $\mathrm{E}_{\boldsymbol{R}}[\rho_{R,G}]$ was termed as "data defect index". Whereas this earlier definition has the virtue of having zero defect for equal-probability sampling, it masks the impact of the population size N. The updated definition via $\mathrm{E}_{\boldsymbol{R}}[\rho_{R,G}^2]$ resolves this problem, and it connects directly with MSE.

[4]An excellent question raised by Dr. Alex Adamou during the 2016 Royal Statistical Society (RSS) conference is how to define d.d.i. when our estimand is population maximum or minimum. In general, defining appropriate d.d.i. for estimators and estimands other than sample and population averages is currently an open problem.

2.3. *Understanding d.d.i.: Data defect index.* Among the three terms defined in (I)–(III), $D_O = (1-f)/f$ and $D_U = \sigma_G^2$ are functions of traditionally familiar measures, and their magnitudes are well understood: both can vary freely on $[0, \infty)$ with no mathematical constraints between them in general. In comparison, the d.d.i. $D_I = \mathrm{E}_{\boldsymbol{R}}[\rho_{R,G}^2]$ is new. As such, readers may have (at least) three questions:

(A) What are the likely magnitudes of D_I, when we have probabilistic samples?
(B) How do we calculate or estimate D_I for non-probabilistic data?
(C) Theoretically, can D_I take any value in $[0, 1]$, for a given D_O and D_U?

To address question (A), let us consider the most basic building block in probabilistic sampling, the simple random sampling (SRS). Under SRS, $\overline{G}_n$ is unbiased for $\overline{G}_N$ and its mean squared error (MSE) is the same as its variance:

$$\mathrm{V}_{\mathrm{SRS}}(\overline{G}_n) = \frac{1-f}{n} S_G^2 \qquad \text{with } S_G^2 = \frac{N}{N-1}\sigma_G^2, \tag{2.5}$$

where $(1-f)$ is known as the "finite sample correction" [e.g., Kish (1965)]. Substituting (2.5) to the left-hand side of (2.4) immediately reveals that the d.d.i. for any SRS is given by

$$D_I \equiv \mathrm{E}_{\mathrm{SRS}}\big(\rho_{R,G}^2\big) = \frac{1}{N-1}. \tag{2.6}$$

In Section 3, we will show that this $D_I \propto N^{-1}$ phenomenon holds for probabilistic sampling in general, and hence D_I will be vanishingly small for large populations. It is this finding, which will be fully explored in later sections, that provides the critical insight to most troubles for dealing with non-probabilistic Big Data sets when their D_I's do not vanish with N^{-1}.

The question (B) has a short answer: we cannot estimate D_I from the sample itself. By definition, everyone in the sample has $R_J = 1$, and hence there is no direct[5] information in the sample for estimating $\rho_{R,G}$. Logically, this has to be the case because if there were meaningful ways to estimate $\rho_{R,G}$ from the same sample, identity (2.3) would then immediately permit us to estimate the *actual* error $\overline{G}_n - \overline{G}_N$, which is impossible without any knowledge/assumption about the R-mechanism. However, this observation also implies that when we are able to ascertain the actual error, such as post-elections, we can ascertain $\rho_{R,G}$ and hence D_I, as demonstrated in Section 4.2 using a polling dataset from 2016 US presidential election. The hope here is that because $\rho_{R,G}$ and hence D_I captures individual response behaviors, there are patterns and lessons to be learned that can help to generate more reliable prior information for future elections. More

[5]Indirect information can exist because of the mathematical constraints imposed by the known sampling rate f and marginal information about G, as shall be discussed shortly.

generally, by borrowing information from similar datasets (e.g., from historical or neighboring studies), we may be able to construct a reasonable prior for $\rho_{R,G}$ or D_I, which minimally would permit us to conduct an informative sensitivity study. For example, the state-wise election data from the 2016 US presidential election allow us to form histograms of $\rho_{R,G}$ (see Section 4.2), which can be used as a plausible prior distribution of $\rho_{R,G}$ for the 2020 US presidential election.

For question (C), the answer is more involved, because $\rho_{R,G}$ is determined by the joint distribution of $\{R_J, G_J\}$ induced by the uniform distribution over J, but D_O and D_U are characteristics of the marginal distributions of R_J and of G_J, respectively. Although marginal means and variances (e.g., f, σ_G^2) are not often perceived to affect correlations (e.g., $\rho_{R,G}$), in general they do impose restrictions because of the Hoeffding identity [Höffding (1940)]

$$\operatorname{Cov}(X, Y) = \int\int \big[F_{X,Y}(x, y) - F_X(x)F_Y(y)\big]\,dx\,dy \tag{2.7}$$

and the Fréchet bounds [Fréchet (1951)]

$$\max\big\{F_X(x) + F_Y(y) - 1, 0\big\} \le F_{X,Y}(x, y) \le \min\big\{F_X(x), F_Y(y)\big\}, \tag{2.8}$$

where $F_{X,Y}$ is a joint cumulative distribution function (CDF) with F_X and F_Y being its two marginal CDFs. The restriction can be particularly severe (mathematically) with discrete variables, especially binary ones. To see this, suppose G_j is also binary, for example, $G_j = 1$ if the jth person plans to support Donald Trump and $G_j = 0$ otherwise. Let $p_G = P_J(G_J = 1)$ and $O_G = p_G/(1 - p_G)$, that is, the odds for voting for Trump. Then, as a special case of Hoeffding–Fréchet bounds, we have

$$-\min\left\{\sqrt{\frac{D_O}{O_G}}, \sqrt{\frac{O_G}{D_O}}\right\} \le \rho_{R,G} \le \min\left\{\sqrt{O_G D_O}, \frac{1}{\sqrt{O_G D_O}}\right\}, \tag{2.9}$$

where the upper bound is achieved by $R_J = G_J$ (e.g., a person responds to the survey if and only if the person plans to vote for Trump), and the lower bound by $R_J = 1 - G_J$ (e.g., a person responds if and only if the person does not plan to vote for Trump). Figure 1 helps to visualize (2.9) in terms of the restrictions on $\rho_{R,G}$ as imposed by p_G and f, where we see that the restrictions are more severe when either f or p_G becomes extreme, that is, very close to zero or one.

As a numerical illustration, if we take $O_G = 1$, and $D_O = 99$ (e.g., 1% of the voter population responded), then (2.9) yields $|\rho_{R,G}| \le 0.1005$. Whereas such bounds might seem very restrictive, we will see shortly, both from theory (Section 3.1) and from the 2016 US election data (Section 4.2, especially Figure 8), that they are far looser than likely in practice, as otherwise our sample results would be essentially useless. Nevertheless, the existence of these bounds suggests caution when we intuitively consider the "smallness" of $\rho_{R,G}$, or when we set values of D_I for theoretical investigations. We must always check if our choices of

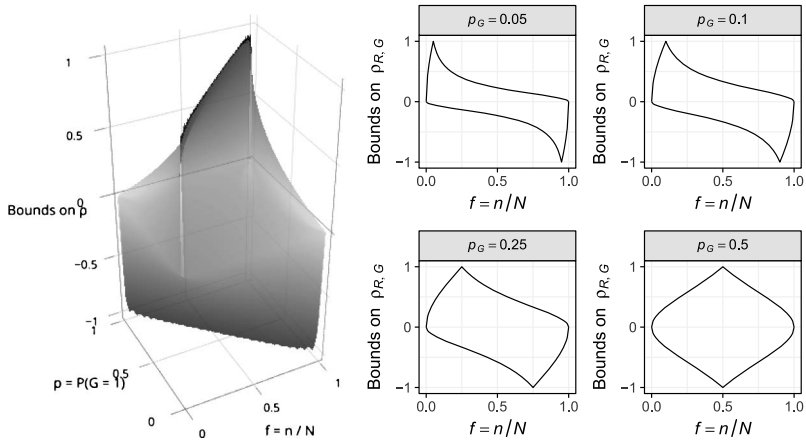

FIG. 1. *The shadowed region in the* 3*D plot depicts the permissible values of* $\{\rho_{R,G}, p_G, f\}$ *as determined by* (2.9). *The restrictions become more severe for* $\rho_{R,G}$ *as* $\{f, p_G\}$ *moves further away from the line* $f = p_G$ *for* $\rho_{R,G} > 0$ *or from the line* $f = 1 - p_G$ *for* $\rho_{R,G} < 0$. *The* 2*D plots show several slices, explicating further how* $\{p_G, f\}$ *restricts* $\rho_{R,G}$.

"*I O U*" correspond to impossible scenarios, or more importantly, to extreme scenarios. For the rest of this article, such a caution is always exercised, and links to more familiar measures (e.g., underreporting probabilities) are presented whenever possible.

3. Compensating for quality with quantity is a doomed game.

3.1. *A law of large populations?* Under a probabilistic sampling, a central driving force for the stochastic behaviors of the sample mean (and alike) is the sample size n. This is the case for the Law of Large Numbers (LLN) and for the Central Limit Theorem (CLT), two pillars of theoretical statistics, and of much of applied statistics because we rely on LLN and CLT to build intuitions, heuristic arguments, or even models. However, identity (2.3) implies that once we lose control of probabilistic sampling, then the driving force behind the estimation error is no longer the sample size n, but rather the population size N. Specifically, (2.3) and (2.5) together imply that

$$Z_{n,N} \equiv \frac{\overline{G}_n - \overline{G}_N}{\sqrt{\mathrm{V}_{\mathrm{SRS}}(\overline{G}_n)}} = \frac{\rho_{R,G}\sqrt{\frac{1-f}{f}}\sigma_G}{\sqrt{\frac{1-f}{n}S_G^2}} = \sqrt{N-1}\rho_{R,G}. \tag{3.1}$$

We emphasize that, although the Z notation is used above to highlight its connection with the usual Z-score, the "Z-score" here is a nominal one because the actual MSE of $\overline{G}_n$ can be very different from its benchmark under SRS, $\mathrm{V}_{\mathrm{SRS}}(\overline{G}_n)$. Indeed, identity (3.1) tells us exactly how they differ, a statistical insight which perhaps deserves to be labeled as a statistical law.

Law of Large Populations (LLP) *Among studies sharing the same (fixed) average data defect correlation* $\mathrm{E}_{\boldsymbol{R}}(\rho_{R,G}) \neq 0$*, the (stochastic) error of* $\overline{G}_n$*, relative to its benchmark under SRS, grows with the population size* N *at the rate of* $\sqrt{N}$.

More precisely, (3.1) tells us that the exact error of the sample mean, as an estimator of the population mean, is $\sqrt{N-1}\rho_{R,G}$ away from zero in the unit of SRS standard error.

The LLP can also be expressed in terms of the so-called *design effect* [Deff; Kish (1965)], or more appropriately the "lack-of-design effect" for non-probabilistic Big Data, which is simply $\mathrm{E}_{\boldsymbol{R}}(Z_{n,N}^2)$:

$$\text{Deff} = \frac{\mathrm{E}_{\boldsymbol{R}}[\overline{G}_n - \overline{G}_N]^2}{\mathrm{V}_{\mathrm{SRS}}(\overline{G}_n)} = (N-1)\mathrm{E}_{\boldsymbol{R}}\big(\rho_{R,G}^2\big) = (N-1)D_I. \tag{3.2}$$

Therefore, under the mean-squared error (MSE) criterion, the design effect of any R-mechanism, with or without any deliberate (probabilistic) design, is exactly $(N-1)D_I$. It is worth noting that, traditionally, the design effect has been defined in terms of variance. But for Big Data, the variance is typically negligible (or even exactly zero if we treat the R-mechanism as deterministic), which makes understanding and assessment of the systematic bias induced by R-mechanism so critical, because it dominates the MSE. Furthermore, like the concept of d.d.i., in general, the effect of sampling design is estimator-dependent. Nevertheless, it is a common practice in the literature to define Deff with respect to sample mean, as the most basic estimator to benchmark the impact of a probabilistic sample design, which controls the data quality. It is therefore natural to adopt the same estimator as we extend the notion of Deff to cover an arbitrary R-mechanism. From a practical perspective, the great algebraic simplicity of (3.1), that is, $Z_{n,N} = \sqrt{N-1}\rho_{R,G}$, also supports the use of sample mean as the benchmarking estimator for data quality, as well as the use of its standard error under SRS as the yardstick unit for comparing estimation errors. The use of SRS yardstick also has a deeper reason. The notion of sample size, *when used as the sole index of information*, is meaningful only when all samples of the same size (from a common population) are probabilistically indistinguishable, that is, when $\Pr(\boldsymbol{R} | \sum_{j=1}^N R_j = n)$ depends only on n and N, which implies SRS.

To state LLP precisely in terms of Deff, let us imagine that we have a sequence of populations with increasing sizes $\{N_\ell, \ell = 1, 2, \ldots\}$ such that $\lim_{\ell\to\infty} N_\ell = \infty$, but with constant sampling rate $f > 0$ (and problem difficulty $D_U = \sigma_G^2$). This induces a sequence of sample sizes $n_\ell = fN_\ell \to \infty$. This setting permits us to use common notation such as $A_N = O(B_N)$ and $A_N = o(B_N)$, which mean, respectively, $\limsup_{\ell\to\infty}(|A_N|/|B_N|) < \infty$ and $\limsup_{\ell\to\infty}(|A_N|/|B_N|) = 0$. With this notation, the identity (3.2) immediately implies that $D_I = O(N^{-1})$ if and only if $\text{Deff} = O(1)$, which is the same as $\mathrm{MSE}_{\boldsymbol{R}}(\overline{G}_n) = O(n^{-1})$ because of (2.5). Consequently, we have

THEOREM 1. *For a fixed sampling rate* $0 < f < 1$ *and problem difficulty* $D_U = \sigma_G^2$, *the following three conditions are equivalent for any R-mechanism*:

(1) *It has a finite design effect*: $\text{Deff} = O(1)$;

(2) *The MSE of the sample mean decreases at the* n^{-1} *rate*: $\text{MSE}_{\boldsymbol{R}}(\overline{G}_n) = O(n^{-1})$;

(3) *Its d.d.i. for the sample mean is controlled at the* N^{-1} *level*: $D_I = O(N^{-1})$.

This result shows explicitly that in order for the sample average to enjoy the usual n^{-1} rate for its MSE, we must control its d.d.i. at the rate of N^{-1}, or equivalently the data defect correlation $\rho_{R,G}$ at the (stochastic) rate of $N^{-1/2}$, regardless of the choice of G (as long as σ_G is finite and fixed). All known probabilistic sampling schemes are designed to achieve condition (1) and equivalently (2) [see Fuller (2011)], and hence the corresponding $D_I = O(N^{-1})$ by Theorem 1, regardless of the sampling rate f or the choice of estimand G. This invariance to f or G is critical for general data quality assurance. It is not difficult to speculate that few haphazard or individually driven R-mechanism can possess such invariance properties unless it is effectively equivalent to a probabilistic sampling, for example, an individual decides to answer or not by flipping a coin (which does not need to be fair, as long as the mechanism of choosing the Jth coin is independent of G_J). For large populations, such as the US eligible voter population, achieving $\rho_{R,G} \approx N^{-1/2}$ for arbitrary sampling rate f without probabilistic sampling (equivalent) requires a miracle. For example, for the 2016 US population of actual voters, $N \approx 1.4 \times 10^8$. To reach $\rho_{R,G} \approx N^{-1/2}$ then requires ensuring $\rho_{R,G} \approx 8.4 \times 10^{-5}$, an extremely small correlation coefficient to be guaranteed from a self-regulated selection mechanism.

Nevertheless, it is worth pointing out that when f is extremely close to one or zero, $D_I = O(N^{-1})$ can be achieved without deliberate probabilistic sampling. For example, for binary G such that $p_G = P_J(G_J = 1) = 1/2$, we see from (2.9) that for any R-mechanism,

$$D_I \le \min\left\{\frac{f}{1-f}, \frac{1-f}{f}\right\} = \min\left\{\frac{n}{N-n}, \frac{N-n}{n}\right\}. \tag{3.3}$$

Therefore, if n_ℓ or $N_\ell - n_\ell$ is bounded by a constant as N_ℓ grows to infinity,[6] then $D_I = O(N^{-1})$. Intuitively, when nearly all R_j's take the same value, either one or zero, then its correlation with any other variable that is not controlled by n cannot be significantly far from zero. But typical situations of Big Data are exactly outside of that "comfort" zone, that is, the sampling rate f is neither close to zero, as in traditional surveys, nor close to one, as in a census.

[6]This phenomenon clearly cannot happen when we assume the sampling rate $f = n_\ell/N_\ell$ is invariant to ℓ.

Some readers might find the symmetry displayed in (3.3) counterintuitive, because it seems to suggest that a sample with size n has the same data defect as a sample with size $N-n$. Shouldn't the latter be far more informative than the former, especially as n is near zero? This symmetry is not a bug, but a feature of d.d.i., because it separates the issue of data quality from the accumulation of information due to data quantity. As far as a selection mechanism goes, selecting n individuals takes the same scheme/effort regardless whether later they are assigned to be respondents or non-respondents. The difference in the amount of information in the resulting datasets is an issue of data quantity, captured by $D_O = (N-n)/n$, no longer symmetric with respect to n and $N-n$. Recall the identity (2.4) says that the difficulty-standardized MSE, $\mathrm{MSE}_{\boldsymbol{R}}(\overline{G}_n)/\sigma_G^2$, is controlled by the product $D_I D_O$. From (3.3), this product is bounded above by $(1-f)^2/f^2$ when $f > 1/2$, but the bound becomes the (rather trivial) constant 1 when $f \le 1/2$. Therefore, for the case underlying (3.3), the product goes to zero only when $f \to 1$, despite the fact that D_I goes to zero whenever $f \to 1$ or $f \to 0$. In the latter case, $f \to 0$ is canceled out in the product by $D_O = (1-f)/f \to \infty$. This fact illustrates once more the importance to consider the tradeoff between data quality and data quantity as captured by the product $D_I D_O$, instead of each term on its own.

3.2. *A butterfly effect*: *The return of the long-forgotten monster* N. To quantify how much damage a seemingly small $\rho_{R,G}$ can cause, we use identity (2.4) to calculate the effective sample size n_{eff} of a Big Data set by equating the MSE of its estimator $\overline{G}_n$ of (2.1) to the MSE of the SRS estimator with the sample size n_{eff}. By (2.4) and (2.5), this yields

$$D_I D_O = \left(\frac{1}{n_{\mathrm{eff}}} - \frac{1}{N}\right)\left(\frac{N}{N-1}\right). \tag{3.4}$$

Let $n_{\mathrm{eff}}^* = (D_O D_I)^{-1}$, then (3.4) implies that

$$n_{\mathrm{eff}} = \frac{n_{\mathrm{eff}}^*}{1 + (n_{\mathrm{eff}}^* - 1)N^{-1}}. \tag{3.5}$$

Under the (trivial) assumption that $n_{\mathrm{eff}}^* \ge 1$, we then have

$$n_{\mathrm{eff}} \le n_{\mathrm{eff}}^* = \frac{f}{1-f} \times \frac{1}{D_I} = \frac{n}{1-f} \times \frac{1}{N D_I}, \tag{3.6}$$

which demonstrates clearly that for probabilistic samples the impact of N on n_{eff}^* (and hence on n_{eff}) is canceled out by D_I because $N D_I = O(1)$, a consequence of Theorem 1. However, once[7] $D_I = O(1)$, however small, $N D_I$ increases with N quickly, leading to a dramatic reduction of n_{eff}.

[7]Mathematically, we need only $D_I = O(N^{\alpha})$ with $\alpha > -1$ in order for $N D_I$ to go to infinity when $N \to \infty$, but whether α can meaningfully take values other than zero (and -1) in practice is an open problem.

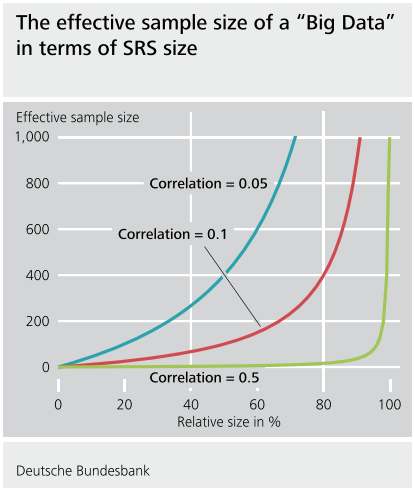

FIG. 2. *Effective sample size* n^*_{eff} *as a function of the relative size* (*sampling rate*) f. [*I thank Dr. Jens Mehrhoff for his permission to use this figure from Mehrhoff* (2016).]

To illustrate, suppose $\mathrm{E}_{\boldsymbol{R}}[\rho_{R,G}] = 0.05$, which seems rather small by conventional standards [e.g., in a similar context of expressing bias, Senn (2007) considered a correlation -0.05 "extremely modest"]. Then $D_I = \mathrm{E}_{\boldsymbol{R}}(\rho^2_{R,G}) \geq [\mathrm{E}_{\boldsymbol{R}}(\rho_{R,G})]^2 = 1/400$. Hence by (3.6), we have

$$n_{\text{eff}} \leq 400\frac{f}{1-f}. \tag{3.7}$$

That is, even if we have data from half of the population, that is, $f = 1/2$, the effective sample size, in terms of an equivalent SRS sample, cannot exceed $n_{\text{eff}} = 400$. But half of the population means about 115 million people for the eligible voter population of US in 2016. Consequently, the "extremely modest" average correlation 0.05 has caused at least a $(115{,}000{,}000 - 400)/115{,}000{,}000 = 99.999965\%$ reduction of the sample size, or equivalently estimation efficiency. The reduction would be even more extreme if we considered the Chinese or India population, precisely because of the impact of the population size. Figure 2, provided by Mehrhoff (2016), visualizes the difficulty of achieving decent effective sample sizes (e.g., between 100–1000) for (average) correlation $\rho_{R,G} = 0.05$, 0.1, and 0.5, albeit one hopes that $\rho_{R,G} \geq 0.1$ is merely a mathematical possibility, permissible by (2.9).

Such dramatic reductions appear to be too extreme to be believable, a common reaction from the audiences whenever this result is presented (see Acknowledgements). It is indeed extreme, but what should be unbelievable is the magical power of probabilistic sampling, which we all have taken for granted for too long. As seen in (3.5), n_{eff} is determined by the product $D_I D_O$. Whereas $D_O = (1-f)/f$ does go to zero when n approaches N, its rate is governed by the relative size f. Clearly it takes a much larger n for $1 - f$ to become negligible than for n^{-1} to become so, and most troublesome of all is that the former depends on the value of

N; $n = 115{,}000{,}000$ makes n^{-1} practically zero for most inference purposes, but it does not make $1 - f$ negligible for almost any problem when $N = 230{,}000{,}000$.

Therefore, the central message here is that once we lose control over the R-mechanism via probabilistic schemes, we can no longer keep the monster N at bay, so to speak. Without the magical power of probabilistic sampling, the right-hand side of (3.2), that is, $(N-1)D_I$, will explode with N. That is, we have a "butterfly effect"—a tiny perturbation caused by D_I can lead to catastrophic error in the end for large N, which in turn causes the seemingly incomprehensible loss of effective sample size. It is therefore essentially wishful thinking to rely on the "bigness" of Big Data to protect us from its questionable quality, especially for large populations.

We remark here that mathematically, it is important to carry the N^{-1} term in the denominator of (3.5), so n_{eff} reaches n for SRS, and $n_{\text{eff}} = N$ instead of $n_{\text{eff}} = \infty$ when $D_I = 0$. For practical purposes, however, it is easier algebraically and crisper conceptually to use n^*_{eff}, which also serves as an almost exact upper bound.[8] We therefore use n^*_{eff} in subsequent calculations. Theoretically, it also provides a simple (almost exact) rule for assessing the impact of the data defect correlation $\rho_{R,G}$: *the effective sample size* n_{eff} *is inverse proportional to* $\rho^2_{R,G}$. For example, if we can cut down $|\rho_{R,G}|$ by 20%, then we will increase n_{eff} by a factor of $(1 - 0.2)^{-2} = 1.5625$, and hence by more than 50%.

Furthermore, in reaching (3.7), we have used the inequality $D_I = \mathrm{E}_{\boldsymbol{R}}(\rho^2_{R,G}) \geq [\mathrm{E}_{\boldsymbol{R}}(\rho_{R,G})]^2$. But the difference between the two sides of this inequality is precisely $\mathrm{V}_{\boldsymbol{R}}(\rho_{R,G})$, which is typically negligible for large N. For example, when the components of $\boldsymbol{R}$ are identically and independently distributed (before conditioning on $\sum_{j=1}^N R_j = n$), then $\mathrm{V}_{\boldsymbol{R}}(\rho_{R,G}) = O(N^{-1})$ (recall this is a conditional variance conditioning on $\sum_{j=1}^N R_j = n$). Hence the variation in $\rho_{R,G}$ caused by random $\boldsymbol{R}$ is negligible compared to D_I precisely when D_I matters, that is, when $\mathrm{E}_{\boldsymbol{R}}(\rho_{R,G})$ does not vanish with N, that is, when $D_I = O(1)$. Consequently, for practical purposes, we usually can ignore the uncertainty in $\rho^2_{R,G}$ as an estimator of its mean, D_I, when N is large and $D_I = O(1)$, as typical with Big Data.

3.3. *A big data paradox?* We statisticians certainly are responsible for the widely held belief that the population size N is not relevant for inference concerning population means and alike, as long as N is sufficiently large. But apparently we have been much less successful in communicating the "warning label" that this assertion is valid only if one has strict control of the sampling scheme (via probabilistic schemes). An effective way to deliver this warning is to observe that the representation (3.1) implies that any routinely used confidence intervals of the form

$$\left(\overline{G}_n - \frac{M\hat{\sigma}_G}{\sqrt{n}}, \overline{G}_n + \frac{M\hat{\sigma}_G}{\sqrt{n}}\right) \tag{3.8}$$

[8] Indeed, it is easy to show that $0 \leq n^*_{\text{eff}} - n_{\text{eff}} < 1$ whenever $(n^*_{\text{eff}} - 1)^2 < N$.

will *almost surely miss* $\overline{G}_N$ for any conventional choice of the multiplier M unless we adopt a $\hat{\sigma}_G$ that overestimates σ_G by orders of magnitude to compensate for the colossal loss of the sample size. Worse, since the interval width in (3.8) shrinks with the apparent size n, our false confidence may increase with n, despite the fact that the interval (3.8) has little chance to cover the truth because it is so precisely centered at a wrong location; its width misses a (huge) factor[9] of $\sqrt{N-n}|\rho_{R,G}|$.

To see this, consider the case of estimating voting preference during the 2016 US presidential election, which will be treated in detail in Section 4. For the current illustration, imagine a researcher has access to self-reported voting preference from 1% of US eligible voter population, whose size is $N \approx 231{,}557{,}000$ [McDonald (2017)]. Let $\hat{p}$ be the sample average from the $n(\approx 2{,}315{,}570)$ observations. Suppose that the uninformed researcher adopts a normal approximation to form a confidence interval for the corresponding population p based on the usual Z-score

$$Z_n = \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} = \frac{\sqrt{n}\sqrt{D_O}\rho_{R,G}}{\sqrt{1-D_O\rho_{R,G}^2-\sqrt{D_O}\rho_{R,G}(\sqrt{\frac{p}{1-p}}-\sqrt{\frac{1-p}{p}})}}, \tag{3.9}$$

where the second expression is obtained by applying the identity (2.3) using the notation $D_O = (1-f)/f$ of (2.4), and with $\sigma_G = \sqrt{p(1-p)}$. We have changed the notation from $Z_{n,N}$ of (3.1) to Z_n here because, following common practice, the latter does not use the finite population correction $(1-f)$, but it does use an estimated $\hat{\sigma}_G^2 = \hat{p}(1-\hat{p})$ instead of $\sigma_G^2 = p(1-p)$. The difference between $Z_{n,N}$ and Z_n is typically inconsequential (as demonstrated below).

Without realizing the self-selected nature of the sample, the uninformed researcher would likely compare Z_n to the $N(0,1)$ reference distribution for constructing his/her confidence interval. The normality is not much of an issue with such a large n, but the mean of Z_n is far away from 0. Consider the case $p = 1/2$ with $\rho_{R,G} = 0.005$, which we shall see in Section 4 is a rather realistic magnitude for Trump supporters' reporting mechanism. Inserting these values together with $D_O = 99$ and $n \approx 2{,}315{,}570$, we obtain

$$Z_n = \sqrt{2{,}315{,}570}\sqrt{\frac{99\times 0.005^2}{1-99\times 0.005^2}} = 75.80. \tag{3.10}$$

Consequently, unless the researcher uses a normal interval with at least "75 sigma" as its half width, which must sound ridiculously ridiculous, the researcher's interval will miss the target. Note that the value of $|Z_n|$ can also be obtained (approximately) as $\sqrt{n/n^*_{\text{eff}}}$, where the effective sample size $n^*_{\text{eff}} \approx [D_O\rho_{R,G}^2]^{-1} =$

[9]This factor would be $\sqrt{N}|\rho_{R,G}|$ if (3.8) includes the finite-sample correction, that is, with $\hat{\sigma}_G$ replaced by $\sqrt{1-f}\hat{\sigma}_G$.

$[99 \times 0.005^2]^{-1} = 404$ by (3.5). Hence $\sqrt{n/n^*_{\text{eff}}} \approx 75.70$, where the inconsequential difference with (3.10) is due to the use of an estimated p instead of the theoretical p in forming the denominator in (3.9). Even if we reduce $\rho_{R,G}$ to 0.001, and hence $n^*_{\text{eff}} \approx 404 \times 25 = 10{,}100$, $|Z_n|$ will still exceed 15, a virtually probability zero event under $Z_n \sim N(0,1)$.

This seemingly striking phenomenon perhaps deserves the label of paradox.

Big Data Paradox *The bigger the data, the surer we fool ourselves.*

In Section 3.5, we will provide additional reasons why this phenomenon is particularly tied to Big Data. The Big Data Paradox is in the same spirit as Simpson's Paradox, a topic of the sequel of this paper [Meng (2018)]. That is, these kinds of statistical phenomena are not paradoxes in mathematical or philosophical senses; indeed mathematically the probability—however indistinguishable from zero—of (3.8) covering the truth can be a complicated function of n or N, depending partially on how the estimated $\hat{\sigma}_G$ is constructed. But they appear to be paradoxical because of our mis-formed or mis-informed intuitions. Here the phrase *Big Data* refers to those big datasets with an uncontrolled (or unknown) R-mechanism. If our big datasets possess the same high quality as those from well designed and executed probabilistic surveys in terms of $\rho_{R,G}$, then we are indeed in paradise in terms of information gathering—nothing beats high quality big data (albeit we may still face the challenges of processing and analyzing large amounts of complex data).

3.4. *"Weight, weight, don't tell me..."* Instead of relying on the mercy of any non-probabilistic process to deliver data quality, whenever we are concerned with an unacceptably large D_I, we should take actions. Reducing bias through weighting is a popular strategy; see Gelman (2007) and many references therein. Unfortunately, whereas weighting often does reduce estimation error, it does not avoid the curse of large N. To see this, let $W_j \geq 0$ be the weight we use for G_j and define $\widetilde{R}_j = R_j W_j$; to simplify notation here we use $\widetilde{A}$ to indicate the weighted version of A. The weighted sample average then is

$$\widetilde{G}_n = \frac{\sum_{j=1}^N R_j W_j G_j}{\sum_{j=1}^N R_j W_j} = \frac{\mathrm{E}_J[\widetilde{R}_J G_J]}{\mathrm{E}_J[\widetilde{R}_J]}. \tag{3.11}$$

We can then generalize (2.2) by replacing the binary R_J with the more general $\widetilde{R}_J$, which leads to

$$\begin{aligned} \widetilde{G}_n - \overline{G}_N &= \frac{\mathrm{Cov}_J(\widetilde{R}_J, G_J)}{\mathrm{E}_J(\widetilde{R}_J)} = \rho_{\widetilde{R},G} \sqrt{\frac{\mathrm{V}_J(\widetilde{R}_J)}{\mathrm{E}_J^2(\widetilde{R}_J)}}\sigma_G \\ &= \rho_{\widetilde{R},G} \times \sqrt{\frac{1 - f + CV_W^2}{f}} \times \sigma_G, \end{aligned} \tag{3.12}$$

where CV_W is the coefficient of variation (i.e., standard deviation/mean) of W_J given $R_J = 1$, that is, among those whose G values are recorded. Historically, Senn (2007) used (3.12) in the case of $f = 1$ to express the difference between weighted individual estimates and their arithmetic mean in the context of meta analysis.

Comparing (3.12) with (2.3), we see the use of weight W affects the actual estimation error in two ways. The negative impact is that it introduces an extra factor

$$a_W = \frac{\sqrt{\frac{1-f+CV_W^2}{f}}}{\sqrt{\frac{1-f}{f}}} = \sqrt{1 + \frac{CV_W^2}{1-f}} \geq 1.$$

Hence, when $|\rho_{R,G}| = |\rho_{\widetilde{R},G}|$, the weighting would necessarily lead to a larger *actual* error in magnitude. This negative impact is particularly pronounced when the (relative) variation in the weights is high, as measured by CV_W^2, and the sampling rate f is high. The deterioration of a weighted estimator as CV_W^2 increases is a well-known fact, in both the survey and importance sampling literature [e.g., see Liu (1996), Owen (2013)], albeit there the increase is generally understood in terms of variance, not the actual error. The fact that higher CV_W^2 will cause relatively more damage to larger datasets can be understood by considering the extreme case when $f = 1$. In such a case, the equally weighted estimator is the population mean and hence it has zero error. Therefore, any error introduced by unequal weighting will render infinite relative error, which is correctly captured by $a_W = \infty$. The only time unequal weighting does not introduce error is when $\sigma_G^2 = 0$, or when we are extremely lucky to produce exactly zero correlation between $\widetilde{R}_J = R_J W_J$ and G_J.

This last point also hints at the goal of using weights. Our hope is that by a judicious choice of W_J, we can reduce the data defect correlation, that is, achieving $|\rho_{\widetilde{R},G}| < |\rho_{R,G}|$, to the degree that this positive impact would outweigh the negative one to ensure that $\widetilde{\text{Deff}} < \text{Deff}$. Here $\widetilde{\text{Deff}}$ is the design effect for $\widetilde{G}_n$ [still with $\text{V}_{\text{SRS}}(\overline{G}_n)$ as the benchmark], which, because of (3.12), is given by

$$\widetilde{\text{Deff}} = \frac{\text{E}_{\boldsymbol{R}}[\widetilde{G}_n - \overline{G}_N]^2}{\text{V}_{\text{SRS}}(\overline{G}_n)} = (N-1)\widetilde{D}_I A_W, \tag{3.13}$$

where $\widetilde{D}_I = \text{E}_{\boldsymbol{R}}(\rho_{\widetilde{R},G}^2)$, and $A_W = \text{E}_{\boldsymbol{R}}[a_W^2] = 1 + (1-f)^{-1}\text{E}_{\boldsymbol{R}}[CV_W^2] \geq 1$. Note $\text{E}_{\boldsymbol{R}}[CV_W^2]$ is used here instead of CV_W^2 because in general the weights themselves may be estimated, and that for $\widetilde{G}_n$, (3.6) is still applicable as long as we replace D_I by $\widetilde{D}_I A_W$ because of (3.13).

The ideal goal, of course, is to bring down $\widetilde{D}_I A_W$ to the level of N^{-1}. But it is rarely possible to do so when the weights themselves are subject to errors, typically much larger than $O(N^{-1})$ for large N [see Kim and Kim (2007), Kim and

Riddles (2012)]. To see this clearly, we write $\pi_j = \Pr_{\boldsymbol{R}}(R_j = 1|\boldsymbol{X})$, which captures the potential bias created by the R-mechanism (recall $\pi_j = \mathbb{1}_{\{R_j=1\}}$ if R_j is deterministic). Note here $\boldsymbol{X} = \{X_1, \ldots, X_N\}$, and hence we permit π_j to be influenced by $X_i, i \neq j$, though often such cross-individual influence can be ignored. If π_j is known for those observed G_j, then a well-known weighting scheme is to set $W_j \propto \pi_j^{-1}$, which leads to the Horvitz–Thompson estimator [Horvitz and Thompson (1952)]. From the perspective of d.d.i., such weighting aims to reduce the mean of $\rho_{\widetilde{R},G}$ to zero by ensuring $\mathrm{E}_{\boldsymbol{R}}[\mathrm{Cov}_J(\widetilde{R}_J, G_J)|\boldsymbol{X}] = 0$, which holds when $W_j \propto \pi_j^{-1}$. However, π_j is never known exactly or even approximately when the R-mechanism is at best partially understood, which is the case for virtually all the observational Big Data sets. Horvitz–Thompson type estimators are known to be extremely sensitive to the errors in the estimated weights because a small $\hat{\pi}_j$ can cause a very large and dominating weight $W_j \propto \hat{\pi}_j^{-1}$. Many methods have been proposed in the literature, such as trimming and power shrinkage [e.g., Chen et al. (2006), Gelman (2007)]. But none of them suggests the remote possibility of reducing $\rho_{\widetilde{R},G}$ to the order of $N^{-1/2}$, especially for large N.

Indeed, Gelman (2007) emphasizes that many weighting schemes go beyond the inverse probability weighting, which introduce additional errors and variations, and hence he opened his article with the line "Survey weighting is a mess". The title of this sub-section[10] echoes Gelman's frustration, even without referencing to the more stringent requirement to deal with the large populations underlying the messier Big Data. But without bringing the d.d.i. down to the level of N^{-1}, we will be destined for surprises if we put our confidence on the sheer size of a Big Data set to compensate for its unknown quality, as illustrated in Section 3.3.

3.5. *Answering the motivating question.* Having defined the d.d.i. D_I, we can give a quantitative answer to the question: "Which one should I trust more: a 1% survey with 60% response rate or a self-reported administrative dataset covering 80% of the population?". Specifically, when we compare MSEs of two sample averages from two datasets for the same variable, identity (2.4) tells us that which estimator is better would depend on neither the quality index D_I nor the quantity index D_O alone, but on their product, providing a precise recipe for tradeoff. To meaningfully answer the motivating question then requires additional information on how the two D_I's compare. To be concrete, suppose our first dataset is a probabilistic sample with sampling rate $f_s = n_s/N$ and design effect Deff. Without non-response, we know $D_I^{(s)} = \mathrm{Deff}/(N-1)$ from (3.2). With non-response, the resulting d.d.i., D_I, is likely to be significantly larger than $D_I^{(s)}$ when there is non-response bias. Furthermore, the sampling rate is reduced to rf_s, where r is the response rate, hence its $D_O = (1 - rf_s)/(rf_s)$.

[10] I thank Professor Doug Rivers, Chief Scientist at YouGov, for this most humorous and telling line, and for email exchanges regarding CCES and elections in general.

Now suppose our second dataset is a Big Data set with data defect index D_I^{BIG} and dropout odds $D_O^{\mathrm{BIG}} = (1-f)/f$. Then by (3.4), its $n_{\mathrm{eff}}^{\mathrm{BIG}}$ is larger than the n_{eff} of the first dataset if and only if

$$D_I^{\mathrm{BIG}} D_O^{\mathrm{BIG}} < D_I D_O. \tag{3.14}$$

To translate this condition into one that can render practical guidelines, we denote the *dropout odds ratio* by

$$\mathcal{O} = \frac{D_O}{D_O^{\mathrm{BIG}}} = \frac{1-rf_s}{rf_s} \times \frac{f}{1-f}. \tag{3.15}$$

Using the approximation $D_I \approx \rho_{R,G}^2$ (see Section 3.2), inequality (3.14) becomes

$$|\rho_{R,G}^{\mathrm{BIG}}| \le \sqrt{\mathcal{O}}|\rho_{R,G}|. \tag{3.16}$$

Condition (3.16) provides us with a practical guideline and a base for sensitive study, even though we typically do not know $\rho_{R,G}^{\mathrm{BIG}}$ or $\rho_{R,G}$. For example, if we are reasonably sure that the mechanism leading to non-response in our survey is similar to the mechanism responsible for self-reporting behavior in the Big Data, then we should be reasonably confident that the Big Data set is more trustworthy when $f \gg rf_s$ because that implies $\sqrt{\mathcal{O}} \gg 1$, and hence (3.16) is very likely to hold. For our question, $f_s = 0.01$, $r = 0.6$, and $f = 0.8$, and hence $\sqrt{\mathcal{O}} \approx 26$, which should be large enough for us to be confident that the 80% administrative data set is more trustworthy.

On the other hand, if we believe that the selection bias caused by the non-response mechanism in the sample is not nearly as severe as in the Big Data set, then we need to have a reasonable sense of the magnitude of $\rho_{R,G}$ before we can become confident that (3.16) holds simply because $\sqrt{\mathcal{O}}$ is large. Our knowledge of the population size is useful for this assessment. Suppose the population underlying our question is the US eligible voter population in 2016. Then $N \approx 231{,}557{,}000$, and hence for SRS, $|\rho_{R,G}^{(s)}| \approx \sqrt{2/\pi}(N-1)^{-1/2} = 5.2 \times 10^{-5}$ [here we use the fact $\mathrm{E}|Z| = \sqrt{2/\pi}$ when $Z \sim N(0,1)$]. Suppose the non-response mechanism has increased the data defect correlation 5 times to $\rho_{R,G} \approx 2.6 \times 10^{-4}$, and hence $26 \times \rho_{R,G} \approx 0.0068$. Whereas on its own a correlation of 0.68% seems so small, in Section 4.2 we will see that it is still larger than all the $\rho_{R,G}$'s observed there. Hence if the $\rho_{R,G}$'s from Section 4.2 are relevant for the current example, then we can still put our faith on the large administrative data. However, if the administrative data in our question covers 50% of the population instead of 80%, then $\sqrt{\mathcal{O}} \approx 13$. Consequently, we will need $|\rho_{R,G}^{\mathrm{BIG}}| < 0.0068/2 = 0.0034$ in order to trust the administrative dataset. This bound is now well within the range of the $\rho_{R,G}$'s observed in Section 4.2, and hence one should no longer be so confident that the big administrative data set covering 50% the population is more trustworthy than the 1% survey with 60% response rate, even if the latter itself suffers from

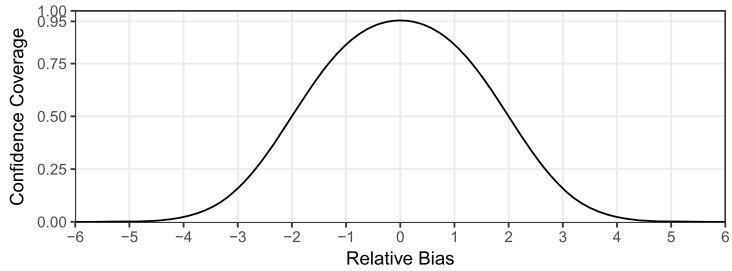

FIG. 3. *Confidence coverage $C(b)$ as a function of the relative bias $b_n = b/\sigma_n$.*

non-response bias. This example demonstrates again the grave consequences of selection bias, because a seemingly trivial data defect correlation can substantially reduce the effective sample size.

One then may ask if we have been fooling ourselves most of the time with survey results (and alike) because almost all of them are subject to non-response biases. However, the issue of non-coverage is not as extreme with small samples as with big datasets because when $D_I = O(1)$, we miss the width of the correct interval by a factor of $\sqrt{n/n_{\text{eff}}}$, which is far more dramatic for Big Data than otherwise. To see this, suppose the mean of $\overline{G}_n$ differs from the estimand $\overline{G}_N$ by an amount of b, and the standard error of $\overline{G}_n$ is σ_n. Then the actual coverage of the usual 95% confidence interval based on the normal approximation, namely, $|\overline{G}_n - \overline{G}_N| < 2\sigma_n$ (we use 2 instead of 1.96 for simplicity), is given (approximately) by $C(b_n) = \Phi(2 - b_n) - \Phi(-2 - b_n)$, where $b_n = b/\sigma_n$, and $\Phi(z)$ is the CDF for $N(0, 1)$. Figure 3 plots $C(b_n)$ against b_n, which shows that as long as $|b_n| < 2$, the coverage will still maintain above 50%. But it deteriorates quickly beyond that, and once $|b_n| > 5$, the coverage becomes essentially zero. Therefore, ironically, the small-sample variance, which helps to reduce the value of b_n because of larger value of σ_n, has provided us with some protections against being completely misled by the selection bias induced by the R-mechanism.

Nevertheless, the concept of d.d.i. and more broadly the issue of data quality is critical for any kind of data, small or large. Its dramatic effect on Big Data population inferences should serve as a clear warning of the serious consequences of ignoring it. The next section demonstrates how to assess d.d.i. in practice, in the context of the 2016 US presidential election, providing a quantitative measure of our collective overconfidence, leading to a big surprise on November 8, 2016.

4. Applications to binary outcome and the 2016 US general election.

4.1. *A measure of overconfidence in* 2016 *US presidential election.* As discussed before, the data defect correlation $\rho_{R,G}$ is not a quantity that has been well studied, partly because it is not directly estimable. Here we use the 2016 US presidential election as a background setting to connect it with the bias in reporting

propensity, a more familiar quantity. We will reveal some simple formulas for assessing how non-response biases affect the effective sample size and hence the margin of error with binary outcomes.

For the 2016 US presidential election, many (major) polls were conducted and reported by media, especially in the last several weeks before the election, as many as about 50 in a single day (see www.realclearpolitics.com/epolls/latest_polls/president/). By a very rough "guesstimate", putting all these polls together amounts to having opinions from up to 1% of the eligible voter population, that is, $f = 0.01$ or $n \approx 2{,}315{,}570$. Any reputable statistician likely would agree that it would be too optimistic to treat the combined poll as a high-quality probabilistic sample with the same size n. But what would be an appropriate discount factor? Cut it to half? By a factor of 10?

To answer this question in the cleanest way, let us assume that there are no complications other than non-response. For example, *response bias* is negligible, as is the percentage of voters who changed their minds over the period when these surveys were taken or the percentage of people appearing in more than one poll. All these complications can only *further* reduce our confidence in the polling results. To calculate D_I, we let $X_j = 1$ if the jth individual plans to vote for Trump and $X_j = 0$ otherwise. Let p_X be the population vote share for Trump, and recall f is the percentage of people who (honestly) report their plans. Then it is easy to verify that the population correlation between X_J and R_J over the uniform distribution on J is given by

$$\rho_{R,X} = \frac{P_J(X_J = 1, R_J = 1) - p_X f}{\sqrt{p_X(1-p_X)}\sqrt{f(1-f)}} = \Delta_R \sqrt{\frac{p_X(1-p_X)}{f(1-f)}}, \tag{4.1}$$

where

$$\Delta_R = P_J\big(R_J = 1 | X_J = 1\big) - P_J\big(R_J = 1 | X_J = 0\big) \tag{4.2}$$

is the reporting bias in absolute terms. Here the notation Δ_R implies that it is determined by the realized R; its expectation, denoted by $\Delta = \mathrm{E}_{\boldsymbol{R}}(\Delta_R)$, would be zero for SRS (assuming no non-responses). Also, by (4.1), whenever $\mathrm{V}_{\boldsymbol{R}}(\rho_{R,X}) = O(N^{-1})$, so would $\mathrm{V}_{\boldsymbol{R}}(\Delta_R)$ (for given p_X and f). Consequently, as before, when Δ matters, that is, when it approaches zero no faster than $O(N^{-1/2})$, typically $O(1)$, we can ignore the differences between Δ_R and Δ or between Δ_R^2 and $\mathrm{E}_{\boldsymbol{R}}(\Delta_R^2)$.

As an illustration, assuming $\Delta = -0.001$ (for reasons to be given in Section 4.2), $f = 0.01$ and $p_X = 0.5$ [Trump's vote share was 0.461 in terms of the popular vote, but $\sqrt{p_X(1-p_X)}$ is very stable near $p_X = 0.5$: $\sqrt{0.461 \times 0.539} = 0.4985$], $\rho_{R,X} = -0.001 \times 0.5/\sqrt{0.01 \times 0.99} = 1/(200\sqrt{0.99}) = -0.00502$. Hence the d.d.i. $D_I = \rho_{R,X}^2 = 1/(200\sqrt{0.99})^2 = 1/39{,}600$. Consequently, by (3.5), the effective sample size of the combined dataset with $n = 2{,}315{,}570$ is

$n^*_{\text{eff}} = \frac{0.01}{0.99} \times 39{,}600 = 400$. Or we can obtain this directly by

$$n^*_{\text{eff}} = \frac{1}{p_X(1-p_X)}\left(\frac{f}{\Delta}\right)^2 = 4\left(\frac{0.01}{0.001}\right)^2 = 400. \tag{4.3}$$

This represents an over 99.98% loss of sample size compared to $n = 2{,}315{,}570$, similar to what we have seen before. We note that because $p_X(1-p_X) \le 1/4$ for any value of $p_X \in [0,1]$, the convenient expression, which requires no knowledge of p_X:

$$\widehat{n}^*_{\text{eff}} = 4\left(\frac{f}{\Delta}\right)^2 \tag{4.4}$$

serves as a lower bound, as well as a very good approximation for the type of election polls where assessing margins of error is particularly important (because p_X's are close to $1/2$). Historically, the fact $p_X(1-p_X) \le 1/4$ has led to an exceedingly simple (and rather accurate) upper bound of the margin of error—denoted by M_e—for SRS[11] proportion with size n_s (ignoring the finite population correction when $n_s \ll N$). That is, if we adopt the usual multiplier 2 for achieving 95% confidence, then the half-width of the 95% confidence interval, namely, M_e, satisfies

$$M_e = 2\sqrt{\frac{p_X(1-p_X)}{n_s}} \le \frac{1}{\sqrt{n_s}}. \tag{4.5}$$

Therefore, an effective sample size of 400 implies that M_e is about 5%, which is 83 times larger than $1/\sqrt{n} \approx 0.06\%$, the margin of error from using the apparent size $n \approx 2{,}315{,}570$. The latter would lead to gross overconfidence in what the data would actually tell us.

An astute reader may point out that $\Delta = -0.1\%$ is not small relatively, because f is only 1%. For example, if the actual $\Delta_R = -0.1\%$ and $p_X = 50\%$, then $P_J(X_J = 1|R_J = 1) = 47.5\%$ and $P_J(X_J = 0|R_J = 1) = 52.5\%$. That is, there are 5% fewer Trump voters in the sample than non-Trump voters, hence a strong bias (in the sample) against Trump given the actual voting preference is a tie. However, even if $\Delta = -0.01\%$, that is, on average we only have 0.5% fewer Trump voters than non-voters in our sample, which would reduce D_I by a factor of 100, we would have $n^*_{\text{eff}} = 40{,}000$ by (4.3), still a 98.27% reduction in the actual sample size compared to $n = 2{,}315{,}570$. The margin of error would be reduced to 0.5%, which still matters for a tight race, unlike when $M_e = 0.06\%$, which becomes negligible. Recall the actual popular vote difference between Clinton and Trump is about 0.8% among eligible voters, or 1.5% among actual voters.

It is worth noting that the above calculations were done using the population of eligible voters, not of the actual voters, because before an election we

[11] The $1/\sqrt{n_s}$ bound on the margin of error is also very useful for other more sophisticated sampling methods such as stratified sampling, because these methods use SRS as a benchmark to improve upon.

do not know who—or even how many—would actually turn out to vote. But as a retrospective investigation, we could repeat the same exercise by assuming our targeted population consists of those who actually turned out to vote, which for the 2016 presidential race was $N^{(a)} \approx 136{,}700{,}730$ [McDonald (2017)]. If we retain the same n based on the (optimistic) assumption that all respondents to the polls cast their votes, then $f^{(a)} \approx 1.7\%$. Consequently, by (4.3) or (4.4), $n^*_{\text{eff}} = 4 \times 17^2 = 1156$ when $\Delta = -0.001$, and $n^*_{\text{eff}} = 115{,}600$ when $\Delta = -0.0001$. These number represent, respectively, a 99.95% and 95% loss of (effective) sample size when compared to $n = 2{,}315{,}570$, resulting in corresponding margin of errors $M_e = 2.9\%$ and $M_e = 0.29\%$, still outside the comfort (confidence) zone indicated by $M_e = 0.06\%$.

4.2. *Estimating d.d.i. for Trump's supporters in CCES surveys.* Because $\rho_{R,G}$ is a dimensionless correlation, designed to measure a responding behavior or recording mechanism, it is not wishful thinking to assume that it can be reasonably assessed from a similar study for which one ultimately is able to assess the actual bias. For example, if we let $\widetilde{B} = \widetilde{G}_n - \overline{G}_N$ be the actual bias from a weighted estimator, then from (3.12), we have

$$\rho_{\widetilde{R},G} = \frac{\widetilde{B}}{\sigma_G \sqrt{(1 - f + CV_W^2)/f}}, \tag{4.6}$$

which reduces to the simpler expression under equal weighting, that is, when $CV_W = 0$,

$$\rho_{R,G} = \frac{\overline{G}_n - \overline{G}_N}{\sigma_G \sqrt{(1-f)/f}} \equiv \frac{B}{\sqrt{D_O D_U}}. \tag{4.7}$$

We emphasize here that which sample estimator to use depends on which responding/recording mechanism we want to assess. If our goal is to assess d.d.i. of the original raw sample, then we must use the unweighed estimator $\overline{G}_n$, as in (4.7), regardless how bad it is as an estimator of $\overline{G}_N$. On the other hand, if we want to assess the imperfection in some weights, that is, how much data defect correlation still remains after applying the weighting, then we should use the corresponding weighted estimator $\widetilde{G}_n$, as in (4.6).

To illustrate the utility of (4.6)–(4.7), we use data from the 2016 Cooperative Congressional Election Study (CCES), a national stratified-sample online survey administered by YouGov. The CCES is considered to be one of the largest and most reliable national election surveys, conducted by a team of leading political scientists [see Ansolabehere, Schaffner and Luks (2017)]. Its individual-level data, available on-line, go back to 2005. The exact stratification specifications, however, are currently unavailable in the database we have used. Ideally we want to assess d.d.i. for all kinds of weighted samples using various survey weights, a task that will be completed by a subsequent project. Here we focus on using equal weights

as a starting point (except for a turnout adjustment used for a comparison), assessing the data defects in the raw sample.

Each year the CCES fields two waves of surveys, a pre-election wave in the weeks before the general election and a post-election wave. Most of its respondents are recruited from YouGov's opt-in online panel, as well as panels of other survey firms. Invitation to respond to the survey originates from a matched sample, which approximates a sampling frame that is representative of the U.S. adult population. In addition to voting preference, the survey asks about general political attitudes, various demographic information, assessment of key policy issues, and political knowledge.

To use this rich data source to assess the actual bias B or more generally $\widetilde{B}$, we need to address the issue of a mis-match between the surveyed population—which at the best is the eligible voter population—and the actual voter population. This mis-match is a well-known issue [Burden (2000)], and there are a number of methods (used by political scientists) to reduce its impact, such as using an estimated propensity for voting as weight or using a subsample of respondents who are validated to be actual voters from public record voter files after the election [Ansolabehere and Hersh (2012)]. No method is fault-proof (e.g., validation via a matching algorithm is subject to errors), an issue that will be explored in detail in a subsequent work. But the general patterns of the findings for our purposes here have been rather robust to the actual method used. The top row of Figure 4 plots the comparisons of the state-wise actual vote shares by Clinton versus the three estimates from CCES data. Whereas the specific estimates have some differences, the overall pattens are very similar: the CCES estimates over-predict in Republican states, under-predict in Democratic states, and are just about right in swing states.[12] The bottom row in Figure 4 provides the counterpart results for Trump, where the general pattern is uniform under-prediction for all states, with the sole exception being Washington D.C., an outlier where Clinton won with over 90% of the vote.

Regardless of which method is used to assess the d.d.i., the resulting estimate is of interest when a similar method would be used in the future for a similar election. This is because, as emphasized in Section 2, d.d.i. aims to capture defects in both data collection and data analysis, including the choice of the estimates. Because the results based on validated voters avoid the issue of weighting, and are likely more reliable (e.g., predicting voting turnout is known to be difficult), we will use them as a simple illustration for estimating the d.d.i. D_I. The state-level results from CCES permit an examination of evidence to contradict the hypothesis that $D_I = O(N^{-1})$, that is, there is no detectable evidence for selective response.

[12]Following the Cook Political Report, swing states (green) are states whose 2016 presidential results flipped from 2012 or whose winner won by less than 5 percentage points. Solidly Republican states (red) are states Romney won in 2012 and which Trump won by more than 5 percentage points. Solidly Democratic states (blue) are states Obama won in 2012 and which Clinton won by more than 5 percentage points.

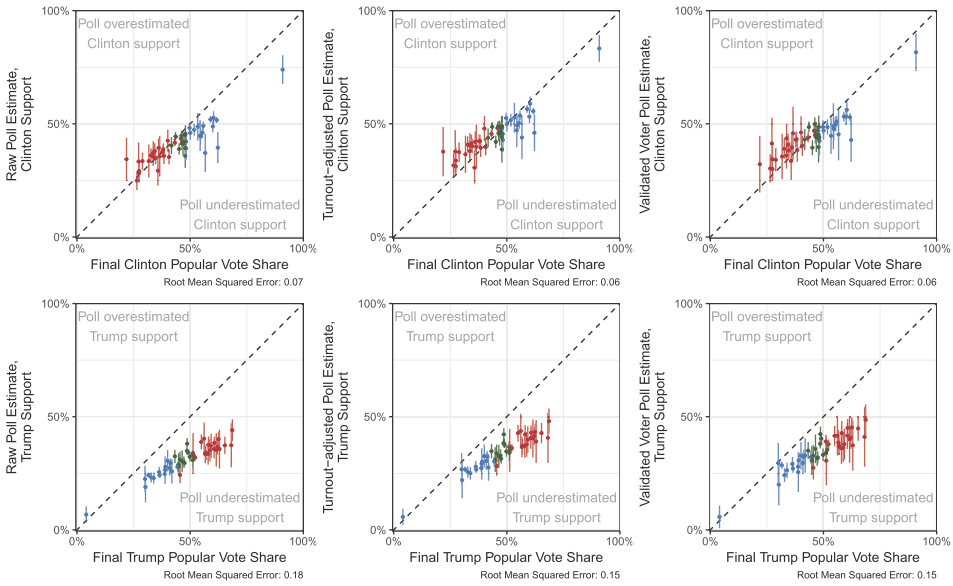

FIG. 4. *Comparison of actual vote shares with CCES estimates (and 95% confidence interval) across 50 states and DC. Top row for Clinton; bottom row for Trump. Color indicates a state's partisan leanings in* 2016 *election: solidly Democratic (blue), solidly Republican (red), or swing state (green). The left plot uses sample averages of the raw data* ($n = 64{,}600$) *as estimates; the middle plot uses estimates weighted to likely voters according to turnout intent (estimated turnout* $\hat{n} = 48{,}106$*); and the right plot uses sample averages among the subsample of validated voters (subsample size* $n = 35{,}829$*). Confidence intervals based on unweighted sample proportions are computed following* (3.9)*, where the use of SRS variances can be conservative given the stratified design of the survey, and yet they still do not provide any realistic protection against the increased MSE caused by the non-response bias. For the turnout adjusted estimate, which is in a ratio form, a* δ*-method is employed to approximate its variance, which is then used to construct confidence intervals.*

Specifically, Figure 5 plots the histograms of the estimated state-level data defect correlation $\hat{\rho}_N$, where we switch the notation from $\rho_{R,G}$ to $\hat{\rho}_N$ to emphasize the (potential) strong dependence of $\rho_{R,G}$ on the population size N, and we use

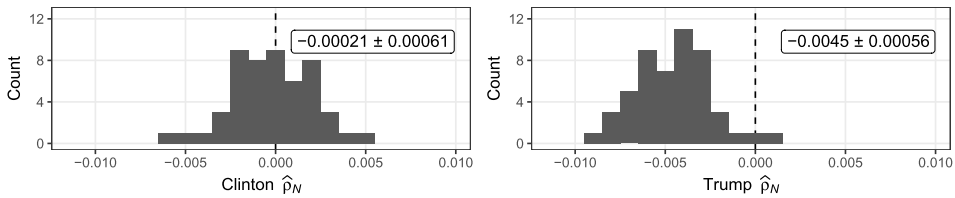

FIG. 5. *Histograms of state-level data defect correlations assessed by using the validated voter data: Clinton's supporters (left) versus Trump's supporters (right). The numbers in boxes show "mean* $\pm$ 2 *standard error".*

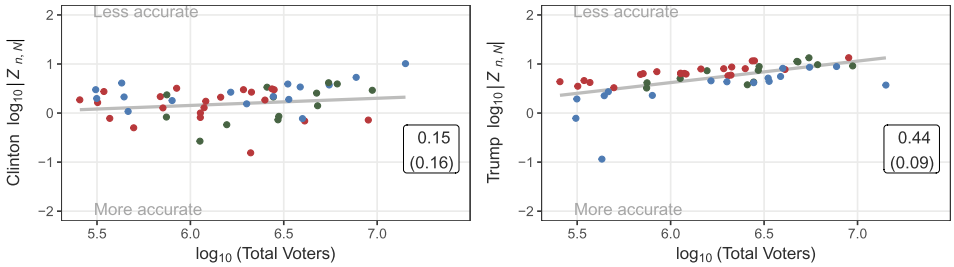

FIG. 6. *Estimates of* $\log|Z_{n,N}| = \hat{\alpha} + \hat{\beta}\log N$: *The numbers in boxes show the least-squares estimate* $\hat{\beta}$ *and its standard error* (*in parentheses*). *States are colored as in Figure* 4.

the "hat" notation to highlight the dependence of $\rho_{R,G}$ on the particular realization of $\boldsymbol{R}$. We will denote $\rho_N = \mathrm{E}_{\boldsymbol{R}}(\hat{\rho}_N)$. We see a distinctive pattern between the histogram for Clinton's supporters (left), which centers around zero, and the one for Trump's (right), which is almost exclusively below zero, and centers quite close to -0.005, a strong indication of higher non-response probability for Trump's supporters. It is important to emphasize that here the response variable is coded as binary, that is, voting for a candidate or not. Because the CCES surveys have an "undecided" category (in addition to third parties), not preferring Trump does not imply voting for Clinton. Otherwise $\hat{\rho}_N$ would be of the same value for Clinton's and Trump's supporters, except with opposite signs. (More generally, we can assess d.d.i. with a categorical X.)

Figure 6 provides further—and more visual—evidence for this distinctive pattern, as well as for underreporting from Trump's supporters. The plot on the log scale was motivated by (3.1), which implies that the logarithm of the magnitude of the relative actual error can be written as

$$\log|Z_{n,N}| = \log|\hat{\rho}_N| + 0.5\log(N-1). \tag{4.8}$$

The central idea here is that, if there is a detectable evidence of selective reporting bias, then the value of $\hat{\rho}_N$ should be relatively stable over a set of populations of different sizes N but sharing the same selective R-mechanism, instead of decreasing with $\log N^{-1/2} = -0.5\log N$. Consequently, $\log|Z_{n,N}|$ should increase with $\log N$ with slope $\beta_0 = 0.5$, as (4.8) would suggest [replacing $\log(N-1)$ by $\log N$ is inconsequential]. In contrast, when there is no selection bias, $\log|Z_{n,N}|$ should not vary much with $\log N$ because the $0.5\log N$ term would be balanced out by $\log|\hat{\rho}_N| \propto -0.5\log N$. Therefore, by fitting the regression

$$\log|Z_{n,N}| = \alpha + \beta\log N, \tag{4.9}$$

we can look for evidence to contradict $\beta = 0$ (zero bias induced by R-mechanism) or $\beta = 0.5$ (detectable bias induced by R-mechanism).

Lo and behold, for Clinton's supporters, the least-squares estimator for β is 0.15 with (estimated) standard error 0.16, so $\beta = 0$—but not $\beta = 0.5$—is well within

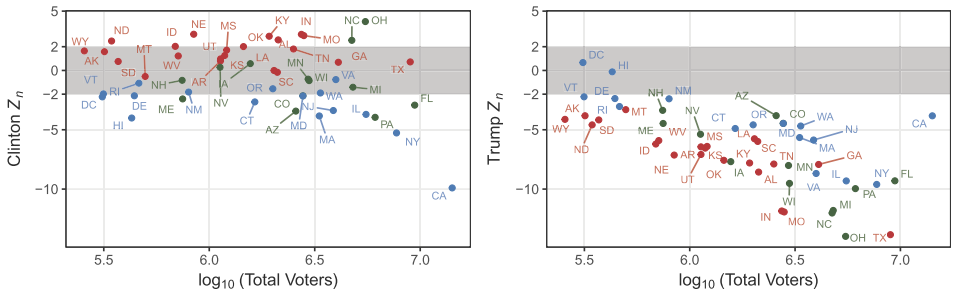

FIG. 7. *Estimates of* $Z_n = \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$: *The conventional* 95% *confidence interval region* $|Z_n| \le 2$ *is indicated in gray.*

the usual margin of error. In contrast, the estimated slope for Trump's supporters is 0.44, quite close to the theoretical value 0.5 in the presence of a biased R-mechanism. Its estimated standard error is 0.09, and hence the usual z-test[13] for testing $\beta = 0$ has a p-value about $p = 10^{-6}$. Using the simple upper bound on the Bayes factor (in favor of the alternative) $B = [-ep\ln(p)]^{-1} \approx 26{,}628$ [Bayarri et al. (2016)], we have very strong odds to reject the null hypothesis of no selective non-response bias.

Figure 7 demonstrates the Big Data Paradox as well as LLP when D_I fails to cancel N^{-1}. For Clinton's vote share, the usual 95% interval based on $|Z_n| \le 2$, where Z_n is given in (3.9), covers 26 out of 51 actual outcomes. This is of course far short of the claimed 95%, a situation which can be significantly improved by using various weights, which we deliberately avoid because we want to assess the actual R-mechanism, as emphasized earlier. But at least the patten of coverage (or lack of) does not indicate a clear trend in terms of the state total turnout, N. In contrast, for Trump's share, there is a very visible monotonic trend with the state-level predictions increasingly moving away from the zone $|Z_n| \le 2$ as N increases. Indeed, Washington DC, Hawaii, and Vermont, the three smallest blue district/states in terms of 2016 turnout, are the only voting regions where the confidence intervals (barely) cover the actual results. This monotonic trend is precisely anticipated by LLP because the error in prediction is proportional to $\sqrt{N}$ whenever the mean of $\hat{\rho}_N$, $\rho_N \neq 0$. And it should provide a clear warning of the Big Data Paradox: it is the larger turnouts that lead to more estimation errors because of systemic (hidden) bias, contrary to our common wisdom of worrying about increased random variations by smaller turnouts. Finally, Figure 8 shows that compared to Clinton's supporters, the data defect correlation $\hat{\rho}_N$ for Trump is closer to its theoretical lower bound given by (2.9), indicating more significant non-response behavior for Trump's voters.

[13]The use of z-test is reasonable given we have 51 data points, and we can assume that under the null, the relative errors are mutually independent.

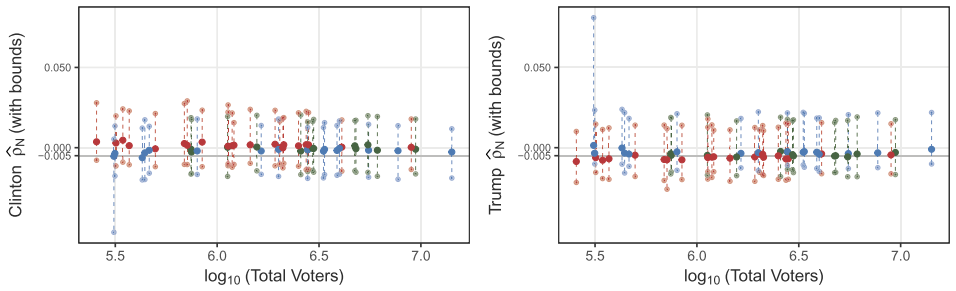

FIG. 8. *Values of* $\hat{\rho}_N$ *with associated theoretical lower and upper bounds from* (2.9). *The* $\hat{\rho}_N = -0.005$ *line approximates the center for Trump's* $\hat{\rho}_N$*'s.*

The above analysis was repeated for various sub-samples, such as by each state's partisan leaning. Whereas smaller samples naturally led to more variable results, the general patterns remain. That is, there is a consistent pattern of underreporting by Trump's supporters, inducing on average about -0.005 data defect correlation, and hence $D_I \approx 2.5 \times 10^{-5}$. This quantitative finding provides numerical evidence to the general belief that there was serious underreporting by Trump's supporters [see e.g., Cohn (2017), Gelman and Azari (2017)]. The quantitative measure of the bias in terms of the data defect correlation $\hat{\rho}_N$ is of value for predicting future elections, such as the 2020 US presidential election. For instance, if the 2020 election is no more volatile or polarizing than the 2016 one, we can use $0.005 \times \sqrt{N}$ as an upper bound for assessing the increased relative prediction error due to selective reporting, where N would be the estimated voter turnout for a state. For example, the turnout in California for 2016 was about 14 million, while for Oklahoma was about 1.4 million. If these numbers stay the same, then for California the bound on the "Z-score" of (3.1) can be as high as $0.005 \times \sqrt{1.4 \times 10^7} \approx 18$, while the same bound should be about 6 for Oklahoma, roughly 1/3 of the bound for California because $1/\sqrt{10} \approx 1/3$. That is, in the worst scenario, our estimation error can go as high as 18 times the benchmarking SRS standard error for California, but only about 6 times for Oklahoma. Of course these bounds are likely to be too extreme, but the very fact that they were reached historically should serve as a powerful reminder that the force of nature is not always within the confidence band of our mental prowess or heartfelt wishes.

In general, we can build models to link d.d.i. with individual demographics, voting district characteristics, or any other covariates that may help to predict voters' response behaviors. The vast amount of 2016 election surveys out there should provide a very rich database to study survey response behaviors in volatile political or social environments. The data from previous elections, such as those collected by CCES since 2005, can also help to assess time trends, if any, in the data defect correlation. Indeed, one may even wish to examine historical cases such as the infamous 1936 *Literary Digest* poll, which had over 2.3 million respondents,

qualified to be a "Big Data" set [Squire (1988)]. A study of this kind using the CCES data and alike is therefore planned.

More generally, when it is not possible to directly assess ρ_N, borrowing information from similar studies is a common strategy in practice, and this can be done using either likelihood methods or Bayesian methods; see Chen et al. (2018) for a recent example in astrostatistics. In particular, because of (2.3), putting a prior on ρ_N amounts to having borrowed prior information on the potential bias. Re-expressing the relative error $Z_{n,N}$ (3.1) in terms of $\hat{\rho}_N$ facilitates this borrowing, because $\hat{\rho}_N$—more precisely, its mean ρ_N—should be more stable across similar studies due to its mean-and-variance standardized nature, just as coefficients of variation tend to be more stable across studies than means or variances alone. Substantively, because ρ_N is intended to capture individual data recording/reporting behavior, it is possible to take advantage of expert knowledge to justify assumptions of homogeneous data defect correlations for similar populations. For example, an analyst could choose to assume that the blue states share the same ρ_N, as do the red states, but the two may differ significantly. Of course, whenever feasible, one should conduct direct studies, such as a behavior assessment or a non-response follow-up survey, to gather data for estimating ρ_N.

5. Epilogue: From Leonhard Euler to Stephen Fienberg.

5.1. *The Euler's identity*: *A statistical counterpart?* Leonhard Euler was a polymath. Formulas and concepts that were named after him are too numerous to list. But none of them is as universally known as Euler's identity:

$$e^{i\pi} + 1 = 0. \tag{5.1}$$

It is often considered to be the most beautiful and mysterious mathematical identity of all time, because it connects the five most fundamental numbers in a deceivingly simple and intriguing way. Among the five, $\{0, 1, e, \pi, i\}$, the most unusual and unexpected one is the imaginary $i = \sqrt{-1}$.

Incidentally, the identity (2.3) connects the five most fundamental quantities/symbols in statistics in an exceedingly simple yet mysterious way (at least at first sight): mean (μ), standard deviation (σ), correlation (ρ), sample size (n), and population size (N):

$$\hat{\mu} - \mu = \rho\sqrt{\frac{N-n}{n}}\sigma, \tag{5.2}$$

where the notation in (2.3) has been rearranged to highlight the five common symbols. Among them, the most unusual and unexpected one is the population size N, for reasons discussed previously.

A key hope of paralleling (5.2) with (5.1) is to raise the awareness of the former for the broader scientific community, because of the pedagogical value it offers

regarding statistical estimation errors. As seen in (2.3), the identity concisely links the three—and only three—determining factors to the statistical error in our estimator, namely (I) data quality, (II) data quantity, and (III) problem difficulty. We can use the identity (5.2) to help our students learn these fundamental factors in a unified way, especially regarding their tradeoffs. We can categorize our general strategies for reducing statistical errors by which factors they target. For example, many probabilistic sampling schemes aim at factor (I), by bringing down the d.d.i. to the level of N^{-1}, as revealed in Section 3.1. Strategies such as stratifications and using covariance adjustments, which require additional input, aim at factor (III), because they reduce the (sub-)population variances, and hence the problem difficulty [Fuller (2011), Kish (1965)].

5.2. *Going beyond a single population.* Just as Euler's identity has various generalizations [e.g., Argentini (2007)], typically less appealing, (5.2) has variations and extensions. As an example of extensions of (2.3)–(2.4) beyond using a single overall sample average, consider the case of stratified sampling/recording with K strata. Let $\overline{G}_{n_k}$ and $\overline{G}_{N_k}$ be respectively the sample and population means of G for stratum k, and $W^{(k)}$ be the stratum weight $k = 1, \ldots, K$. Then the stratified sampling extension of (2.3) for the stratified estimator $\overline{G}_{n,K} = \sum_{k=1}^{K} W^{(k)} \overline{G}_{n_k}$ is

$$\overline{G}_{n,K} - \overline{G}_N \equiv \sum_{k=1}^{K} W^{(k)} (\overline{G}_{n_k} - \overline{G}_{N_k}) = \sum_{k=1}^{K} W^{(k)} \rho_{R,G}^{(k)} \sqrt{\frac{1 - f^{(k)}}{f^{(k)}}} \sigma_G^{(k)}, \tag{5.3}$$

where $\rho_{R,G}^{(k)}$, $f^{(k)}$ and $\sigma_G^{(k)}$ are respectively the counterparts of $\rho_{R,G}$, f and σ_G of (2.3) for the kth stratum. Assuming the sampling/recording schemes in different strata are independent [and indeed they can be of different nature, as in Meng (1993)], the MSE version of (5.3) becomes

$$\begin{aligned} \mathrm{MSE}_{\boldsymbol{R}}(\overline{G}_n) = {} & \sum_{k=1}^{K} [W^{(k)}]^2 D_I^{(k)} D_O^{(k)} D_U^{(k)} \\ & + \sum_{i \neq j} W^{(i)} W^{(j)} \mathrm{E}_{\boldsymbol{R}}(\rho_{R,G}^{(i)}) \mathrm{E}_{\boldsymbol{R}}(\rho_{R,G}^{(j)}) \sqrt{D_O^{(i)} D_O^{(j)}} \sqrt{D_U^{(i)} D_U^{(j)}}, \end{aligned} \tag{5.4}$$

where all D's are the counterparts in (2.4) for the corresponding strata as indicated by their superscripts. There are, however, new cross-products $\mathrm{E}_{\boldsymbol{R}}(\rho_{R,G}^{(i)}) \mathrm{E}_{\boldsymbol{R}}(\rho_{R,G}^{(j)})$ that cannot be ignored precisely because of the (potential) biases in $\overline{G}_{n_i}$ or $\overline{G}_{n_j}$ due to the R-mechanisms in stratum i and j (which can be very different). The non-vanishing second term on the right-hand side of (5.4) reflects the complicated nature of how these strata-wise biases can interact with each other with unpredictable patterns, including the possibility of counterbalancing each other. When the sample average is unbiased within each stratum, (5.4) will resume the simpler

additive "orthogonal form" without the (mathematically) unpleasant second term. The practical impact of this second term, and how to deal with it, is a worthwhile topic for future study particularly because stratified sampling, and more generally multi-stage sampling, is the backbone of much of sample surveys [Fuller (2011), Kish (1965), Lohr (2009)], including CCES.

5.3. *Applications to Monte Carlo and Quasi Monte Carlo* (*MCQMC*). Given the intrinsic links between sample survey and Monte Carlo methods [see e.g., Meng (2005)], it should come as no surprise that identity (5.2) can be transferred and generalized to study Monte Carlo and Quasi-Monte Carlo (MCQMC) methods. Indeed, pioneering work by Hickernell (2018) provides exactly that, and it generalizes the finite-sample version (5.2) to general (super-population) function forms. In a nutshell, when we use (almost) any form of MCQMC methods to estimate, say, an integration $\mu = \int g(\boldsymbol{x})\nu(d\boldsymbol{x})$, we replace the measure ν by a finitely supported measure $\hat{\nu}$ to form $\hat{\mu} = \int g(\boldsymbol{x})\hat{\nu}(d\boldsymbol{x})$. [See Kong et al. (2003) and Kong et al. (2007) for a general likelihood theory about reformulating Monte Carlo integrations as an estimation of measure ν.] Consequently,

$$\begin{aligned}\hat{\mu} - \mu &= \int g(\boldsymbol{x})(\hat{\nu} - \nu)(d\boldsymbol{x}) = \frac{\langle g, \hat{\nu} - \nu\rangle}{\|g\| \times \|\hat{\nu} - \nu\|} \times \|\hat{\nu} - \nu\| \times \|g\| \\ &\equiv \mathrm{CNF}(g, \hat{\nu} - \nu) \times \mathrm{DSC}(\hat{\nu} - \nu) \times \mathrm{VAR}(g),\end{aligned} \tag{5.5}$$

where $\langle g, h\rangle$ is an inner product on a suitable function space (e.g., a Hilbert space with reproducing kernel K) containing g, and $\|h\| = \sqrt{\langle h, h\rangle}$ is its induced norm (e.g., L^2 norm).

Here, following the terminology of Hickernell (2018), $\mathrm{VAR}(g)$ measures the *variation* of g, and it is the counterpart of σ_G (and hence it is on the standard deviation scale, despite the unfortunate clash of abbreviations between variance and variation). Clearly, the higher the variation of g, the greater the difficulty in estimating its integral. $\mathrm{DSC}(\hat{\nu} - \nu)$ is the *discrepancy* between the estimated measure $\hat{\nu}$ and the original measure ν, and its behavior is controlled by the data quantity in the case of probabilistic sampling. Therefore, it is a counterpart of $\sqrt{(1-f)/f}$, though more generally it also depends on the data locations (e.g., as with Quasi-Monte Carlo) and hence also the dimension of $\boldsymbol{x}$. And finally, $\mathrm{CNF}(g, \hat{\nu} - \nu)$ measures the *confounding* between the integrand g and the error of approximation $\hat{\nu} - \nu$. Such a measure plays a key role because if the confounding is high, that is, the error in approximation $\hat{\nu} - \nu$ is larger at locations where g tends to be larger (in magnitude), then we should expect a higher MCQMC error. Hence $\mathrm{CNF}(g, \hat{\nu} - \nu)$ plays the same role as the data defect correlation $\rho_{R,G}$. See Hickernell (2018) for a full exploration of variations of (5.5), including a Bayesian version, many examples and theory, and lessons learned from examining the "trio" of factors, especially CNF, which had been largely ignored in the MCQMC literature. For example, $|\mathrm{CNF}|$ was replaced by its upper bound 1 in arriving at the classic Koksma–Hlawka inequality [Hickernell (2006)]: $|\hat{\mu} - \mu| \le \mathrm{DSC}(\hat{\nu} - \nu) \times \mathrm{VAR}(g)$, an immediate consequence of (5.5).

5.4. *A Fienberg's dream*: *Increasing data quality and privacy simultaneously?* Stephen Fienberg was a polystat. He had over 600 publications in almost every (reputable) statistical journal, and in many others that usually are not on statisticians' mind. The titles of these non-statistical journals read almost as an alphabetical showcase: from *Accounting Reviews*, *Behavior Science*, *Contemporary Jewry*, to *Journal of Interdisciplinary History*, *Jurimetrics*, *Kybernetik*, and to *Neurotoxicology and Teratology*, *Primates*, *Journal of Zoology*. His research contributions and interests cover a wide spectrum: classical topics that every student in statistics should learn [e.g., categorical data analysis, as in Fienberg (2007)]; emerging fields that even the most knowledgeable statisticians might have trouble describing [e.g., algebraic statistics, as in Fienberg, Petrović and Rinaldo (2011)]; long-lasting issues that most people have an opinion on [e.g., US decennial census, as detailed in Anderson and Fienberg (1999)]; and largely overlooked areas of mostly unrealized importance [e.g., the use of statistics in academic administration, as articulated in Fienberg (1996)].

The issue of data quality, the focus of the current paper, is deeply reflected in many of Steve's papers, ranging from the quality of the US census, to the quality of evidence in court, and to ensure both data quality and confidentiality. In particular, Steve was a co-founder of the *Journal of Privacy and Confidentiality* in 2006, and served as its Editor-in-Chief from 2010 to essentially the end of his life.[14] Steve's writing on data confidentiality started long before it became a hot research topic, and his first substantial paper on this topic appears to be "Conflicts between the needs for access to statistical information and demands for confidentiality" [Fienberg (1994)]. Its abstract, quoted below in its entirety, demonstrates Steve's anticipation of the arrival of the Big Data era, and how the tradeoff between data information and confidentiality would become a pressing issue.

> "With the growth of computer-based government records and the continued collection of statistical data for research, especially in the social sciences, there has been a concomitant growth in the desire to access statistical information by government, industry, and university-based researchers. Moreover, as a result of modern computer technology and ever-expanding computer networks, the costs of data acquisition and transfer continue to drop, and the desirability of access to statistical information collected by others increases. While government statistical agencies and survey researchers have always been concerned about the need to preserve the confidentiality of respondents to ensure the quality of statistical data, these concerns have been heightened by the decline in response rates for censuses and surveys over the past two decades. This paper examines the seeming conflicts between the two perspectives of data access and confidentiality protection and briefly outlines some of the issues involved from the perspectives of governments, statistical agencies, other large-scale gatherers of data, and individual researchers."

Since then, Steve (co-)authored over 40 papers on data confidentiality and privacy, and one of the recurrent themes is the emphasis on the tradeoff between sta-

[14]It is even more remarkable that during this period, Steve also served as a Senior Editor (2010–2012) and then Editor in Chief (2013–2015) of the *Annals of Applied Statistics*, and simultaneously the Editor of the *Annual Review of Statistics and its Application* (2011–2015).

tistical information and data privacy, also known as the utility-risk tradeoff [e.g., Duncan and Fienberg (1997), Fienberg, Rinaldo and Yang (2010)]. The tradeoff here concerns two somewhat competing aspects. On one hand, as Steve emphasized in the quoted abstract, preserving data confidentiality is about ensuring data quality, because the risk of disclosure would greatly discourage people from responding honestly if at all. On the other hand, protecting confidentiality of the data already collected typically means that we need to mask or even distort various information in the data, and hence it would lower the data quality. As with the rest of us, Steve shared the dream of preserving privacy without sacrificing information. But Steve would likely be excited by a larger dream, that is, to increase the data quality while enhancing their privacy. The identities (2.2)–(2.3) suggest that might not be a day dream, because we can introduce "noise" to data to both increase data confidentiality and to decrease the data defect correlation, and hence the d.d.i. to a level that also compensates for the increase in the variance, that is, problem difficulty.

5.5. *The possibility of reducing d.d.i. while enhancing data privacy.* As a proof of concept, consider the situation explored in Fienberg (2010), entitled "The relevance or irrelevance of weights for confidentiality and statistical analyses". Steve's main concern is that while weights are inevitable in practice, disclosing weights themselves could jeopardize data confidentiality, especially when the variations in the weights are large (see Section 3.4). This is because those individuals receiving extreme weights, either large or small, may have higher risk of being identified even if their original data are de-identified. Let us assume that if we disclose only the original data $\{X_j, j \in I_n\}$, the user would have no choice but to use the sample average $\overline{X}_n$ for estimating the population mean $\overline{X}_N$, which will be a biased estimator because $\pi_j = \Pr_{\boldsymbol{R}}(R_j = 1|X_j)$, the *propensity* of responding/recording, depends on X_j. As we emphasized before, here R captures the entire "Recording" mechanism, and hence it takes into account non-response bias, among other possible harmful selection biases. The bottom-line is that if we have used weight $W_j \propto \pi_j^{-1}$, we will have an (approximately) unbiased estimator for $\overline{X}_N$.

Before we proceed, we want to highlight a dual interpretation/representation of $\rho_{R,G}$ that will be useful for the construction below. Specifically, the original formulation of d.d.i., as given in Meng (2014), was in terms of $\rho_{\pi,G}$, the correlation between π_J and G_J. Whereas with $N \to \infty$, $\rho_{R,G}$ and $\rho_{\pi,G}$ will become the same, the former enjoys exactness in the sense that (2.3)–(2.4) involve no approximation when we use $\rho_{R,G}$, and hence the updated version reported in this paper. However, if we model π_j via the logistic regression $\text{logit}(\pi_j) = \alpha + \beta G_j$, then it is known that the maximum likelihood estimate $\hat{\theta} = \{\hat{\alpha}, \hat{\beta}\}$ based on $\{G_j, R_j, j = 1, \ldots, N\}$

must satisfy

$$\begin{aligned} &\sum_{j=1}^{N} R_j = \sum_{j=1}^{N} \hat{\pi}_j \quad \text{and} \\ &\sum_{j=1}^{N} R_j G_j = \sum_{j=1}^{N} \hat{\pi}_j G_j \qquad \text{where } \hat{\pi}_j = \frac{e^{\hat{\alpha}+\hat{\beta}G_j}}{1+e^{\hat{\alpha}+\hat{\beta}G_j}}. \end{aligned} \tag{5.6}$$

We emphasize here that this fitting process is a thought experiment, because in reality we do not have access to all $\{G_j, j = 1, \ldots, N\}$ (though we typically know all R_j).[15] Identity (5.6) is a case of "internal bias calibration" studied in detail by Firth and Bennett (1998), which then implies

$$\overline{G}_n - \overline{G}_N = \frac{\mathrm{Cov}_J(\hat{\pi}_J, G_J)}{\mathrm{E}_J(\hat{\pi}_J)} \quad \text{and} \quad \rho_{\hat{\pi},G} = \rho_{R,G}. \tag{5.7}$$

Consequently, we do not necessarily need to invoke a large-N approximation to connect $\rho_{\pi,G}$, the *expected* data defect correlation, with $\rho_{R,G}$, the *realized* data defect correlation, but rather through the familiar logistic model for recording/reporting propensity. This resembles the situation of the expected Fisher information versus the observed Fisher information, which have related but distinctive meanings; which one should be used is not a matter without controversy [see e.g., Efron and Hinkley (1978)]. The same can be said here because $\rho_{R,G}$ captures the actual estimation error for the data at hand, whereas $\rho_{\pi,G}$ the expected error, that is, bias in $\overline{G}_n$. Which one should we use would depend on how individualized we want our inference to be, as formulated and investigated in Liu and Meng (2016).[16] Nevertheless, (5.7) reminds us that we can de-correlate $\rho_{R,G}$ by de-correlating $\rho_{\hat{\pi},G}$, even for small samples, as long as we can assume the logistic model is reasonable.

Specifically, as a thought experiment, let us assume that the data collector have access to $\boldsymbol{\pi} = \{\pi_j, j = 1, \ldots, N\}$, which they do not wish to release. However, their confidentiality concerns do not rule out the possibility of releasing privacy enhanced data say in the form of a function of X_j and π_j: $X_j^* = h(X_j, \pi_j)$, which would be harder for de-identification, especially when the function h is not disclosed. Identity (2.2) together with its dual representation (5.7) suggests that we should seek a function h such that

$$\begin{aligned} &\text{(A)} \quad \mathrm{E}_J\big(X_J^*\big) = \mathrm{E}_J(X_J) \quad \text{and} \\ &\text{(B)} \quad \big|\mathrm{Cov}_J\big(\pi_J, X_J^*\big)\big| < \big|\mathrm{Cov}_J(\pi_J, X_J)\big|. \end{aligned} \tag{5.8}$$

[15] We can also improve the model fitness by extending $\mathrm{logit}(\pi_j) = \alpha + \beta G_j$ to $\mathrm{logit}(\pi_j) = \alpha + \beta G_j + H(G_j; \theta)$ for a suitable non-linear function of G_j with parameter θ distinct from β. This extension will not alter (5.6) except for the obvious change to the expression of $\hat{\pi}_j$ to include $H(G_j; \hat{\theta})$.

[16] I thank my wonderful (former) student and co-author Keli Liu for reminding me of (5.6)–(5.7).

Here for simplicity of demonstrating possibilities, we have assumed X is univariate; the mathematical construction for the more realistic multivariate X, where the risk of breaching confidentiality is typically higher, can be carried out in a component-wise fashion.

In (5.8), the *mean preserving* requirement (A) ensures that the simple sample average will be consistent as $n \to N$, and the *covariance reduction* requirement (B) aims to reduce the bias in the sample average caused by the R-mechanism. Because the two covariances share the same π_J, it is easy to see that (B) is equivalent to $|\rho_{\hat{\pi},X^*}\sigma_{X^*}| \le |\rho_{\hat{\pi},X}\sigma_X|$. Requirement (B) thus ensures that when we reduce the data defect correlation, we do not unduly inflate the difficulty of the problem introduced by the privacy enhanced variable X^*.

Together, requirements (A) and (B) suggest that we should try a regression-type adjustment in the form of $X_j^* = X_j - \beta(\pi_j - \overline{\pi}_N)$, where $\overline{\pi}_N = \mathrm{E}_J(\pi_J)$ and β is to be determined. The covariance reduction requirement then becomes

$$\big|\mathrm{Cov}_J(\pi_J, X_J) - \beta \mathrm{V}_J(\pi_J)\big| < \big|\mathrm{Cov}_J(\pi_J, X_J)\big|. \tag{5.9}$$

Denoting $\beta_{X,\pi} = \mathrm{Cov}_J(\pi_J, X_J)/\mathrm{V}_J(\pi_J)$ to be the *population* regression coefficient of regressing X_j on π_j (as determined by least-squares), we can re-express (5.9) as

$$|\beta_{X,\pi} - \beta| \le |\beta_{X,\pi}| \quad \Longleftrightarrow \quad -|\beta_{X,\pi}| + \beta_{X,\pi} \le \beta \le |\beta_{X,\pi}| + \beta_{X,\pi}. \tag{5.10}$$

The ideal choice $\beta = \beta_{X,\pi}$ is not achievable in practice because $\beta_{X,\pi}$ is unknown, but it can be approximated by the data collector via regressing X_j on π_j in the sample $j \in I_n$, denoted by $\widehat{\beta}_{X,\pi}$. Requirement (5.10) then is a very good piece of news because it says that as long as we get the sign of $\beta_{X,\pi}$ correct, and do not incur a relative approximation error over 100%, that is, as long as

$$\left|\frac{\hat{\beta}_{X,\pi} - \beta_{X,\pi}}{\beta_{X,\pi}}\right| < 1 \tag{5.11}$$

the user's sample average based on the privacy enhanced data $\{X_j^*, j \in I_n\}$ would have (asymptotically) a smaller MSE than the one based on the original data $\{X_j, j \in I_n\}$. The flexibility provided by (5.11) is the same as that provided by Lemma 1 of Meng (2005) in the context of covariance adjustment for MCQMC estimators, where the question was how large the error in estimation of the regression coefficient needs to be before the covariance (incorrectly) adjusted estimator incurs larger MCQMC error than the unadjusted one. Indeed, one may argue that even if we have the optimal β we may still prefer to use a different one from the permissible range as given in (5.10), because a sub-optimal β in that region may provide more privacy protection than the optimal β since the former would be harder to discover.

Of course a fundamental problem of enhancing confidentiality via adding zero-mean noise is that it preserves population averages only for linear (in data) estimands. How to reliably estimate the weights is another thorny issue as we discussed in Section 3.4. Nevertheless, the construction above reminds of us that

when the data quality has room for improvement, with additional confidential information (such as weights), it is possible to improve the data quality as an integrated part of improving data confidentiality. Such simultaneous improvements are obviously not possible when the data quality is already at the highest level (which seldom happens in reality, if ever). For example, the regression adjustment method above will be of no help when $\beta_{X,\pi}$ is already zero, that is, when $\rho_N = 0$, as with SRS or any equal probability sampling. Incidentally, the above derivation provides another justification of using weights as a covariate [e.g., Gelman (2007)], instead of using them to form the unstable Horvitz–Thompson estimate, as in Section 3.4.

5.6. *A more challenging problem*: *Individualized predictions*. So far we have focused on the issue of (big) data quality for population inferences. However, much of the current excitement generated by Big Data is about the pursuit of *individualization*: from personalized medicine to individualized education to targeted marketing. This desire poses a more challenging problem at the statistical foundational level. Because each of us is unique, any attempt to "personalize" must be approximative in nature. But this is a very different kind of notion of approximation, compared to the traditional large-sample asymptotics, where the setup is to use a sample of individuals to learn about the population they belong to. In contrast, individualized prediction is about finding a sequence of proxy populations with increased resolutions to learn about an individual. This leads to an ultimate challenge for Statistics (and statisticians): *how to build a meaningful theoretical foundation for inference and prediction without any direct data?*

The sequel of the current paper, Meng (2018), will investigate this issue by using another idea explored in Meng (2014), the *multi-resolution framework* borrowed from the wavelets literature, where we will see again the fundamental tradeoff between data quantity and data quality. We can increase the amount of indirect data by matching less on the characteristics that define the target individual, that is, by decreasing the resolution of matching when formulating its proxy populations. But this will decrease the data quality because the resulting proxy populations will be less relevant for the individual and hence their results are likely to be more biased for the target individualized prediction. On the other hand, we can increase the matching resolution and hence obtain more relevant proxy data, but that will necessarily decrease the data quantity and hence increase the variance. Big Data certainly can help to reduce the problem, but they do not escape from this fundamental tradeoff. The paradise for fundamental research is therefore wide open.

Indeed, from a research perspective, what is big about Big Data is the number of intellectually and technologically challenging problems that keep many of us sleepless either because we are too excited or too frustrated. Therefore, the statistical issues touched upon in this article and in the sequel are tiny ice chips at the tip of an iceberg. The literature on Big Data and more generally data science is too vast for any single paper to summarize adequately, but Donoho (2017) and its discussions are definitely a great place to start. A main purpose of this paper

and its sequel is to encourage more (young) talents to enter the emerging paradises of foundational research for Big Data and Data Science, where there is so much learned, being learned, and waiting to be learned.

Acknowledgements. The article was developed for the 2015 Statistician of the Year presentation at ASA (America Statistical Association) Chicago Chapter, an AMS-MAA (America Mathematical Society—Mathematical Association of America) joint address at 2016 JMM (Joint Mathematical Meetings), a plenary address at 2016 International Conferences of RSS (Royal Statistical Society) and of IISA (International Indian Statistical Association), Institute for Advanced Study Distinguished Lecture at Hong Kong University of Science and Technology, a banquet speech for the 10th ICSA (International Chinese Statistical Association) International Symposium, a Public Lecture celebrating the 50th anniversary of University of Calgary, the 2017 Hogg–Craig Lecture at University of Iowa, the 2017 Bahadur Lecture at The University of Chicago, a Late Breaking session on the 2016 elections at 2017 JSM (Joint Statistical Meetings), a 45/45 Lecture of Harvard Data Science Initiative, and the 2018 Arnoff–Schloss Lecture in Lindner College of Business, University of Cincinnati. It was also presented at Arizona State University, University of Pittsburgh, University of Virginia, University of Michigan, McGill University, Rice University, Cambridge University, Colby College, and Oxford University. I thank many audience members there, especially Alex Adamou, Fred Hickernell, Jae-Kwang Kim, Jens Mehrhoff, Stephen Senn, and Naisyin Wang for constructive comments; my many colleagues and students, especially Joseph Blitzstein, Stephen Blyth, Alex Blocker, Richard Born, Yang Chen, David Firth, Alan Garber, Andrew Gelman, Robert Gibbons, Robin Gong, Kosuke Imai, Gary King, Xinran Li, Keli Liu, Thomas Louis, and Steve Stigler for encouraging and enabling me to think bigger and deeper; and the US National Science Foundation and John Templeton Foundation for partial financial support. I am particularly in debt to Jeremy Wu of the US Census Bureau for getting me into this line of research; to my colleague Stephen Ansolabehere of Harvard Government Department for introducing the CCES data and his student and co-author Shiro Kuriwaki, who did an excellent job in carrying out an extensive analysis and producing many figures, some of which are presented in this paper, and hence my heartfelt thanks to Shiro. I am deeply grateful to Todd Kuffner and Kai Zhang for serving as *previewers*, providing extremely helpful comments and independent technical verifications; to editors and referees for very insightful and constructive suggestions that greatly improve the flow of the presentation; and to Steven Finch for careful proofreading. Any defects in the paper are purely mine (even though I don't want them!).

REFERENCES

ANDERSON, M. and FIENBERG, S. E. (1999). *Who Counts? The Politics of Census-Taking in Contemporary America*. Russell Sage Foundation.

ANSOLABEHERE, S. and HERSH, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Polit. Anal.* **20** 437–459.

ANSOLABEHERE, S., SCHAFFNER, B. F. and LUKS, S. (2017). Guide to the 2016 Cooperative Congressional Election Survey. Available at http://dx.doi.org/10.7910/DVN/GDF6Z0.

ARGENTINI, G. (2007). A matrix generalization of Euler identity $e^{ix} = \cos(x) + i\sin(x)$. Preprint. Available at arXiv:math/0703448.

BAYARRI, M. J., BENJAMIN, D. J., BERGER, J. O. and SELLKE, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *J. Math. Psych.* **72** 90–103. MR3506028

BETHLEHEM, J. (2009). The rise of survey sampling. *CBS Discussion Paper* **9015**.

BURDEN, B. C. (2000). Voter turnout and the national election studies. *Polit. Anal.* **8** 389–398.

CHEN, C., DUAN, N., MENG, X.-L. and ALEGRIA, M. (2006). Power-shrinkage and trimming: Two ways to mitigate excessive weights. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* 2839–2846.

CHEN, Y., MENG, X.-L., WANG, X., VAN DYK, D. A., MARSHALL, H. L. and KASHYAP, V. L. (2018). Calibration concordance for astronomical instruments via multiplicative shrinkage. *J. Amer. Statist. Assoc.* To appear.

COHN, N. (2017). Election review: Why crucial state polls turned out to be wrong. *The New York Times*, *June* 1*st*.

DONOHO, D. (2017). 50 years of data science. *J. Comput. Graph. Statist.* **26** 745–766. MR3765335

DUNCAN, G. T. and FIENBERG, S. E. (1997). Obtaining information while preserving privacy: A Markov perturbation method for tabular data. In *Joint Statistical Meetings* 351–362.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65** 457–487. MR0521817

FIENBERG, S. E. (1994). Conflicts between the needs for access to statistical information and demands for confidentiality. *J. Off. Stat.* **10** 115–132.

FIENBERG, S. E. (1996). Applying statistical concepts and approaches in academic administration. In *Education in a Research University* 65–82. Stanford Univ. Press, Stanford.

FIENBERG, S. E. (2007). *The Analysis of Cross-Classified Categorical Data*, Springer Science & Business Media.

FIENBERG, S. E. (2010). The relevance or irrelevance of weights for confidentiality and statistical analyses. *Journal of Privacy and Confidentiality* **1** 183–195.

FIENBERG, S. E., PETROVIĆ, S. and RINALDO, A. (2011). Algebraic statistics for p_1 random graph models: Markov bases and their uses. In *Looking Back. Lect. Notes Stat. Proc.* **202** 21–38. Springer, New York. MR2856692

FIENBERG, S. E., RINALDO, A. and YANG, X. (2010). Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *International Conference on Privacy in Statistical Databases* 187–199. Springer, Berlin.

FIRTH, D. and BENNETT, K. E. (1998). Robust models in probability sampling (with discussions). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 3–21. MR1625672

FRÉCHET, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon. Sect. A.* (3) **14** 53–77. MR0049518

FULLER, W. A. (2011). *Sampling Statistics* Wiley, New York.

GELMAN, A. (2007). Struggles with survey weighting and regression modeling (with discussions). *Statist. Sci.* **22** 153–188.

GELMAN, A. and AZARI, J. (2017). 19 things we learned from the 2016 election (with discussions). *Statistics and Public Policy* **4** 1–10.

HARTLEY, H. O. and ROSS, A. (1954). Unbiased ratio estimators. *Nature* **174** 270–271.

HEITJAN, D. F. and RUBIN, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *J. Amer. Statist. Assoc.* **85** 304–314.

HICKERNELL, F. J. (2006). *Koksma–Hlawka Inequality*. Wiley Online Library.
HICKERNELL, F. J. (2018). The trio identity for Quasi-Monte Carlo error analysis. In *Monte Carlo and Quasi Monte Carlo* (P. Glynn and A. Owen, eds.) 13–37. Springer.
HÖFFDING, W. (1940). Masstabinvariante Korrelationstheorie. *Schr. Math. Inst. u. Inst. Angew. Math. Univ. Berlin* **5** 181–233. MR0004426
HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460
KEIDING, N. and LOUIS, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys (with discussions). *J. Roy. Statist. Soc. Ser. A* **179** 319–376. MR3461587
KIM, J. K. and KIM, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canad. J. Statist.* **35** 501–514. MR2381396
KIM, J. K. and RIDDLES, M. K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling. *Surv. Methodol.* **38** 157.
KISH, L. (1965). *Survey Sampling*. Wiley, New York.
KONG, A., MCCULLAGH, P., MENG, X.-L., NICOLAE, D. and TAN, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussions). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 585–618. MR1998624
KONG, A., MCCULLAGH, P., MENG, X.-L. and NICOLAE, D. L. (2007). Further explorations of likelihood theory for Monte Carlo integration. In *Advances in Statistical Modeling and Inference*. *Ser. Biostat.* **3** 563–592. World Sci. Publ., Hackensack, NJ. MR2416134
LIU, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Stat. Comput.* **6** 113–119.
LIU, K. and MENG, X.-L. (2016). There is individualized treatment. Why not individualized inference? *The Annual Review of Statistics and Its Applications* **3** 79–111.
LIU, J., MENG, X.-L., CHEN, C. and ALEGRIA, M. (2013). Statistics can lie but can also correct for lies: Reducing response bias in NLAAS via Bayesian imputation. *Stat. Interface* **6** 387–398. MR3105229
LOHR, S. L. (2009). *Sampling*: *Design and Analysis*. Nelson Education.
MCDONALD, M. P. (2017). 2016 November general election turnout rates. Available at http://www.electproject.org/2016g.
MEHRHOFF, J. (2016). Executive summary: Meng, X.-L. (2014), "A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it)". Conference handout.
MENG, X.-L. (1993). On the absolute bias ratio of ratio estimators. *Statist. Probab. Lett.* **18** 345–348. MR1247444
MENG, X.-L. (2005). Comment: Computation, survey and inference. *Statist. Sci.* **20** 21–28.
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* (X. Lin et al., eds.) 537–562. CRC Press.
MENG, X.-L. (2018). Statistical paradises and paradoxes in big data (II): Multi-resolution inference, Simpson's paradox, and individualized treatments. Preprint.
OWEN, A. B. (2013). Monte Carlo Theory, Methods and Examples. Available at http://statweb.stanford.edu/~owen/mc/.
ROYALL, R. (1968). An old approach to finite population sampling theory. *J. Amer. Statist. Assoc.* **63** 1269–1279.
RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196
SENN, S. (2007). Trying to be precise about vagueness. *Stat. Med.* **26** 1417–1430. MR2359149
SHIRANI-MEHR, H., ROTHSCHILD, D., GOEL, S. and GELMAN, A. (2018). Disentangling bias and variance in election polls. Unpublished manuscript. Available at http://www.stat.columbia.edu/~gelman/research/unpublished/polling-errors.pdf.
SQUIRE, P. (1988). Why the 1936 literary digest poll failed. *Public Opin. Q.* **52** 125–133.

TROXEL, A. B., MA, G. and HEITJAN, D. F. (2004). An index of local sensitivity to nonignorability. *Statist. Sinica* **14** 1221–1237. MR2126350

DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: meng@stat.harvard.edu