A Trio of Inference Problems That Could Win You a Nobel Prize in Statistics (If You Help Fund It)

45

A trio of inference problems that could win you a Nobel Prize in statistics (if you help fund it)

Xiao-Li Meng

Department of Statistics
Harvard University, Cambridge, MA

Statistical inference is a field full of problems whose solutions require the same intellectual force needed to win a Nobel Prize in other scientific fields. Multi-resolution inference is the oldest of the trio. But emerging applications such as individualized medicine have challenged us to the limit: Infer estimands with resolution levels that far exceed those of any feasible estimator. Multi-phase inference is another reality because (big) data are almost never collected, processed, and analyzed in a single phase. The newest of the trio is multi-source inference, which aims to extract information in data coming from very different sources, some of which were never intended for inference purposes. All of these challenges call for an expanded paradigm with greater emphases on qualitative consistency and relative optimality than do our current inference paradigms.

45.1 Nobel Prize? Why not COPSS?

The title of my article is designed to grab attention. But why Nobel Prize (NP)? Wouldn't it be more fitting, for a volume celebrating the 50th anniversary of COPSS, to entitle it "A Trio of Inference Problems That Could Win You a COPSS Award (and you don't even have to fund it)?" Indeed, some media and individuals have even claimed that the COPSS Presidents' Award is the NP in Statistics, just as they consider the Fields Medal to be the NP in Mathematics.

No matter how our egos might wish such a claim to be true, let us face the reality. There is no NP in statistics, and worse, the general public does not

seem to appreciate statistics as a "rocket science" field. Or as a recent blog (August 14, 2013) in *Simply Statistics* put it: "Statistics/statisticians need better marketing" because (among other reasons)

> "Our top awards don't get the press they do in other fields. The Nobel Prize announcements are an international event. There is always speculation/intense interest in who will win. There is similar interest around the Fields Medal in mathematics. But the top award in statistics, the COPSS award, doesn't get nearly the attention it should. Part of the reason is lack of funding (the Fields is $15K, the COPSS is $1K). But part of the reason is that we, as statisticians, don't announce it, share it, speculate about it, tell our friends about it, etc. The prestige of these awards can have a big impact on the visibility of a field."

The fact that there is more public interest in the Fields than in COPSS should make most statisticians pause. No right mind would downplay the centrality of mathematics in scientific and societal advancement throughout human history. Statistics seems to be starting to enjoy a similar reputation as being at the core of such endeavors as we move deeper into the digital age. However, the attention around top mathematical awards such as the Fields Medal has hardly been about their direct or even indirect impact on everyday life, in sharp contrast to our emphasis on the practicality of our profession. Rather, these awards arouse media and public interest by featuring how ingenious the awardees are and how difficult the problems they solved, much like how conquering Everest bestows admiration not because the admirers care or even know much about Everest itself but because it represents the ultimate physical feat. In this sense, the biggest winner of the Fields Medal is mathematics itself: Enticing the brightest talent to seek the ultimate intellectual challenges.

And *that* is the point I want to reflect upon. Have we statisticians adequately conveyed to the media and general public the depth and complexity of our beloved subject, in addition to its utility? Have we tried to demonstrate that the field of statistics has problems (e.g., modeling ignorance) that are as intellectually challenging as the Goldbach conjecture or Riemann Hypothesis, and arguably even more so because our problems cannot be formulated by mathematics alone? In our effort to make statistics as simple as possible for general users, have we also emphasized adequately that reading a couple of stat books or taking a couple of stat courses does not qualify one to teach statistics?

In recent years I have written about making statistics as easy to learn as possible. But my emphasis (Meng, 2009b) has been that we must make a tremendous collective effort to change the perception that "Statistics is easy to teach, but hard (and boring) to learn" to a reality of "Statistics is hard to teach, but easy (and fun) to learn." Statistics is hard to teach because it is intellectually a very demanding subject, and to teach it well requires both depth in theory and breadth in application. It is easy and fun to learn because

it is directly rooted in everyday life (when it is conveyed as such) and it builds upon many common logics, not because it lacks challenging problems or deep theory.

Therefore, the invocation of NP in the title is meant to remind ourselves that we can also attract the best minds to statistics by demonstrating how intellectually demanding it is. As a local example, my colleague Joe Blitzstein turned our Stat110 from an enrollment of about 80 to over 480 by making it both more real-life rooted and more intellectually demanding. The course has become a Harvard sensation, to the point that when our students' newspaper advises freshmen "how to make 20% effort and receive 80% grade," it explicitly states that Stat110 is an exception and should be taken regardless of the effort required. And of course the NPs in the natural and social sciences are aimed at work with enormous depth, profound impact, and ideally both. The trio of inference problems described below share these features — their solutions require developing some of the deepest theory in inference, and their impacts are immeasurable because of their ubiquity in quantitative scientific inquiries.

The target readership of this article can best be described by a Chinese proverb: "Newborn calves are unafraid of tigers," meaning those young talents who are particularly curious and courageous in their intellectual pursuits. I surely hope that future COPSS (if not NP) winners are among them.

45.2 Multi-resolution inference

To borrow an engineering term, a central task of statistical inference is to separate signal from noise in the data. But what is signal and what is noise? Traditionally, we teach this separation by writing down a regression model, typically linear,

$$Y = \sum_{i=0}^{p} \beta_i X_i + \epsilon,$$

with the regression function $\sum_{i=0}^{p} \beta_i X_i$ as signal, and ϵ as noise. Soon we teach that the real meaning of ϵ is anything that is not captured by our designated "signal," and hence the "noise" ϵ could still contain, in real terms, signals of interest or that should be of interest.

This seemingly obvious point reminds us that the concepts of signal and noise are relative — noise for one study can be signal for another, and vice versa. This relativity is particularly clear for those who are familiar with multi-resolution methods in engineering and applied mathematics, such as wavelets (see Daubechies, 1992; Meyer, 1993), where we use wavelet coefficients below or at a primary resolution for estimating signals. The higher frequency ones are treated as noise and used for variance estimation; see Donoho and Johnstone (1994), Donoho et al. (1995) and Nason (2002). Therefore what counts for

signal or noise depends entirely on our choice of the primary resolution. The multi-resolution framework described below is indeed inspired by my learning of wavelets and related multi-resolution methods (Bouman et al., 2005, 2007; Lee and Meng, 2005; Hirakawa and Meng, 2006), and motivated by the need to deal with Big Data, where the complexity of emerging questions has forced us to go diving for perceived signals in what would have been discarded as noise merely a decade ago.

But how much of the signal that our inference machine recovers will be robust to the assumptions we make (e.g., via likelihood, prior, estimating equations, etc.) and how much will wash out as noise with the ebb and flow of our assumptions? Such a question arose when I was asked to help analyze a large national survey on health, where the investigator was interested in studying men over 55 years old who had immigrated to the US from a particular country, among other such "subpopulation analyses." You may wonder what is so special about wanting such an analysis. Well, nothing really, except that there was not a single man in the dataset who fit the description! I was therefore brought in to deal with the problem because the investigator had learned that I could perform the magic of multiple imputation. (Imagine how much data collection resource could have been saved if I could multiply impute myself!)

Surely I could (and did) build some hierarchical model to "borrow information," as is typical for small area estimations; see Gelman et al. (2003) and Rao (2005). In the dataset, there were men over 55, men who immigrated from that country, and even men over 55 who immigrated from a neighboring country. That is, although we had no direct data from the subpopulation of interest, we had plenty of indirect data from related populations, however defined. But how confident should I be that whatever my hierarchical machine produces is reproducible by someone who actually has direct data from the target subpopulation?

Of course you may ask why did the investigator want to study a subpopulation with no direct data whatsoever? The answer turned out to be rather simple and logical. Just like we statisticians want to work on topics that are new and/or challenging, (social) scientists want to do the same. They are much less interested in repeating well-established results for large populations than in making headway on subpopulations that are difficult to study. And what could be more difficult than studying a subpopulation with no data? Indeed, political scientists and others routinely face the problem of empty cells in contingency tables; see Gelman and Little (1997) and Lax and Phillips (2009).

If you think this sounds rhetorical or even cynical, consider the rapidly increasing interest in individualized medicine. If I am sick and given a choice of treatments, the central question to me is which treatment has the best chance to cure me, not some randomly selected 'representative' person. There is no logical difference between this desire and the aforementioned investigator's desire to study a subpopulation with no observations. The clinical trials testing these treatments surely did not include a subject replicating my description

exactly, but this does not stop me from desiring individualized treatments. The grand challenge therefore is how to infer an estimand with granularity or resolution that (far) exceeds what can be estimated directly from the data, i.e., we run out of enough sample replications (way) before reaching the desired resolution level.

45.2.1 Resolution via filtration and decomposition

To quantify the role of resolution for inference, consider an outcome variable Y living on the same probability space as an information filtration $\{\mathcal{F}_r, r = 0, \ldots, R\}$. For example, $\mathcal{F}_r = \sigma(X_0, \ldots, X_r)$, the σ-field generated by covariates $\{X_0, \ldots, X_r\}$, which perhaps is the most common practical situation. The discussion below is general, as long as $\mathcal{F}_{r-1} \subset \mathcal{F}_r, r = 1, \ldots, R$, where r can be viewed as an index of resolution. Intuitively, we can view $\mathcal{F}_r$ as a set of specifications that restrict our target population — the increased specification/information as captured by $\mathcal{F}_r$ allows us to zoom into more specific subpopulations; here we assume $\mathcal{F}_0$ is the trivial zero-information filter, i.e., X_0 represents the constant intercept term, and $\mathcal{F}_R$ is the maximal filter, e.g., with infinite resolution to identify a unique individual, and R can be infinite. Let

$$\mu_r = \mathrm{E}(Y|\mathcal{F}_r) \quad \text{and} \quad \sigma_r^2 = \mathrm{var}(Y|\mathcal{F}_r)$$

be the conditional mean (i.e., regression) and conditional variance (or covariance) of Y given $\mathcal{F}_r$, respectively. When $\mathcal{F}_r$ is generated by $\{X_0, \ldots, X_r\}$, we have the familiar $\mu_r = \mathrm{E}(Y|X_0, \ldots, X_r)$ and $\sigma_r^2 = \mathrm{var}(Y|X_0, \ldots, X_r)$.

Applying the familiar EVE law

$$\mathrm{var}(Y|\mathcal{F}_r) = \mathrm{E}\{\mathrm{var}(Y|\mathcal{F}_s)|\mathcal{F}_r\} + \mathrm{var}\{\mathrm{E}(Y|\mathcal{F}_s)|\mathcal{F}_r\},$$

where $s > r$, we obtain the conditional ANOVA decomposition

$$\sigma_r^2 = \mathrm{E}(\sigma_s^2|\mathcal{F}_r) + \mathrm{E}\{(\mu_s - \mu_r)^2|\mathcal{F}_r\}. \tag{45.1}$$

This key identity reveals that the (conditional) variance at resolution r is the sum of an estimated variance and an estimated (squared) bias. In particular, we use the information in $\mathcal{F}_r$ (and our model assumptions) to estimate the variance at the higher resolution s and to estimate the squared bias incurred from using μ_r to proxy for μ_s. This perspective stresses that σ_r^2 is itself also an estimator, in fact our best guess at the reproducibility of our indirect data inference at resolution r by someone with direct data at resolution s.

This dual role of being simultaneously an estimand (of a lower resolution estimator) and an estimator (of a higher resolution estimand) is the essence of the multi-resolution formulation, unifying the concepts of variance and bias, and of model estimation and model selection. Specifically, when we set up a model with the signal part at a particular resolution r (e.g., $r = p$ for the linear model), we consider μ_r to be an acceptable estimate for any μ_s with $s > r$. That is, even though the difference between μ_s and μ_r reflects systematic variation, we purposely re-classify it as a component of random variation.

In the strictest sense, bias results whenever real information remains in the residual variation (e.g., the ϵ term in the linear model). However, statisticians have chosen to further categorize bias in this strict sense depending on whether it occurs above or below/at the resolution level r. When the information in the residual variation resides in resolutions higher than r then we use the term "variance" for the price of failing to include that information. When the residual information resides in resolutions lower than or at r, then we keep the designation "bias." This categorization, just as the mathematician's O notation, serves many useful purposes, but we should not forget that it is ultimately artificial.

This point is most clear when we apply (45.1) in a telescopic fashion (by first making $s = r + 1$ and then summing over r) and when $R = \infty$:

$$\sigma_r^2 = \mathrm{E}(\sigma_\infty^2 | \mathcal{F}_r) + \sum_{i=r}^{\infty} \mathrm{E}\{(\mu_{i+1} - \mu_i)^2 | \mathcal{F}_r\}. \tag{45.2}$$

The use of $R = \infty$ is a mathematical idealization of the situations where our specifications can go on indefinitely, such as with individualized medicine, where we have height, weight, age, gender, race, education, habit, all sorts of medical test results, family history, genetic compositions, environmental factors, etc. That is, we switch from the hopeless $n = 1$ (i.e., a single individual) case to the hopeful $R = \infty$ scenario. The σ_∞^2 term captures the variation of the population at infinite resolution. Whether σ_∞^2 should be set to zero or not reflects whether we believe the world is fundamentally stochastic or appears to be stochastic because of our human limitation in learning every mechanism responsible for variations, as captured by $\mathcal{F}_\infty$. In that sense σ_∞^2 can be viewed as the intrinsic variance with respect to a given filtration. Everything else in the variance at resolution r are merely biases (e.g., from using μ_i to estimate μ_{i+1}) accumulated at higher resolutions.

45.2.2 Resolution model estimation and selection

When $\sigma_\infty^2 = 0$, the infinite-resolution setup essentially is the same as a potential outcome model (Rubin, 2005), because the resulting population is of size one and hence comparisons on treatment effects must be counterfactual. This is exactly the right causal question for individualized treatments: What would be my (health, test) outcome if I receive one treatment versus another? In order to estimate such an effect, however, we must lower the resolution to a finite and often small degree, making it possible to estimate average treatment effects, by averaging over a population that permits some degrees of replication. We then hope that the attributes (i.e., predictors) left in the "noise" will not contain enough real signals to alter our quantitative results, as compared to if we had enough data to model those attributes as signals, to a degree that would change our qualitative conclusions, such as choosing one treatment versus another.

That is, when we do not have enough (direct) data to estimate μ_R, we first choose a $\mathcal{F}_{\tilde{r}}$, and then estimate μ_R by $\hat{\mu}_{\tilde{r}}$. The "double decoration" notation $\hat{\mu}_{\tilde{r}}$ highlights two kinds of error:

$$\hat{\mu}_{\tilde{r}} - \mu_R = (\hat{\mu}_{\tilde{r}} - \mu_{\tilde{r}}) + (\mu_{\tilde{r}} - \mu_R). \tag{45.3}$$

The first parenthesized term in (45.3) represents the usual model estimation error (for the given $\tilde{r}$), and hence the usual "hat" notation. The second is the bias induced by the resolution discrepancy between our actual estimand and intended estimand, which represents the often forgotten model selection error. As such, we use the more ambiguous "tilde" notation $\tilde{r}$, because its construction cannot be based on data alone, and it is not an estimator of R (e.g., we hope $\tilde{r} \ll R$).

Determining $\tilde{r}$, as a model selection problem, then inherits the usual bias-variance trade-off issue. Therefore, any attempt to find an "automated" way to determine $\tilde{r}$ would be as disappointing as those aimed at automated procedures for optimal bias-variance trade-off (see Meng, 2009a; Blitzstein and Meng, 2010). Consequently, we must make assumptions in order to proceed. Here the hope is that the resolution formulation can provide alternative or even better ways to pose assumptions suitable for quantifying the trade-off in practice and for combating other thorny issues, such as nuisance parameters. In particular, if we consider the filtration $\{\mathcal{F}_r, r = 0, 1, \ldots\}$ as a cumulative "information basis," then the choice of $\tilde{r}$ essentially is in the same spirit as finding a sparse representation in wavelets, for which there is a large literature; see, e.g., Donoho and Elad (2003), Poggio and Girosi (1998), and Yang et al. (2009). Here, though, it is more appropriate to label $\mu_{\tilde{r}}$ as a parsimonious representation of μ_R.

As usual, we can impose assumptions via prior specifications (or penalty for penalized likelihood). For example, we can impose a prior on the model complexity $\tilde{R}_\delta$, the smallest (fixed) r such that $\mathrm{E}\{(\mu_r - \mu_R)^2\} \le \delta$, where δ represents the acceptable trade-off between granularity and model complexity (e.g., involving more X's) and the associated data and computational cost. Clearly $\tilde{R}_\delta$ always exists but it may be the case that $\tilde{R}_\delta = R$, which means that no lower-resolution approximation is acceptable for the given δ.

Directly posing a prior for $\tilde{R}_\delta$ is similar to using L_0-regularization (Lin et al., 2010). Its usefulness depends on whether we can expect all $X_r's$ to be more or less exchangeable in terms of their predictive power. Otherwise, the resolution framework reminds us to consider putting a prior on the ordering of the X_i's (in terms of predictive power). Conditional on the ordering, we impose priors on the predictive power of incremental complexity, $\Delta_r = \mu_{r+1} - \mu_r$. These priors should reflect our expectation for Δ_r^2 to decay with r, such as imposing $\mathrm{E}(\Delta_r^2) > \mathrm{E}(\Delta_{r+1}^2)$. If monotonicity seems too strong an assumption, we could first break the X_i's into groups, assume exchangeability within each group, and then order the groups according to predictive power. That is to say, finding a complete ordering of the X_i's may require prior knowledge that is too refined. We weaken this knowledge requirement by seeking only an ordering

over equivalence classes of the X_i's where each equivalence class represents a set of variables which we are not able to a priori distinguish with respect to predictive power. The telescoping additivity in (45.2) implies that imposing a prior on the magnitude of Δ_r will induce a control over the "total resolution bias" (TRB)

$$\mathrm{E}(\mu_{\tilde{R}_\delta} - \mu_R)^2 = \sum_{r=\tilde{R}_\delta}^{R} \mathrm{E}(\mu_r - \mu_{r+1})^2,$$

which holds because Δ_r and Δ_s are orthogonal (i.e., uncorrelated) when $s \neq r$.

A good illustration of this rationale is provided when $\mathcal{F}_r$ is generated by a series of binary variables $\{X_0, \ldots, X_r\}$, $r = 0, \ldots, R$. In such cases, our multi-resolution setup is equivalent to assuming a weighted binary tree model with total depth R; see Knuth (1997) and Garey (1974). Here each node is represented by a realization of $\vec{X}_r = (X_0, \ldots, X_r)$, $\vec{x}_r = (x_0, \ldots, x_r)$, at which the weights of its two (forward) branches are given by $w_{\vec{x}_r}(x) = \mathrm{E}(Y|\vec{X}_r = \vec{x}_r, X_{r+1} = x)$ respectively with $x = 0, 1$. It is then easy to show that

$$\mathrm{E}(\Delta_r^2) \leq \frac{1}{4}\,\mathrm{E}\{w_{\vec{X}_r}(1) - w_{\vec{X}_r}(0)\}^2 \equiv \frac{1}{4}\,\mathrm{E}\{D^2(\vec{X}_r)\},$$

where $D^2(\vec{X}_r)$ is a measure of the predictive power of X_{r+1} that is not already contained in $\vec{X}_r$. For the previous linear regression, $D^2(\vec{X}_r) = \beta_{r+1}^2$. Thus putting a prior on $D^2(\vec{X}_r)$ can be viewed as a generalization of putting a prior on the regression coefficient, as routinely done in Bayesian variable selection; see Mitchell and Beauchamp (1988) and George and McCulloch (1997).

It is worthwhile to emphasize that Bayesian methods, or at least the idea of introducing assumptions on Δ_r's, seems inevitable. This is because "pure" data-driven type of methods, such as cross-validation (Arlot and Celisse, 2010), are unlikely to be fruitful here — the basic motivation of a multi-resolution framework is the lack of sufficient replications at high resolutions (unless we impose non-testable exchangeability assumptions to justify synthetic replications, but then we are just being Bayesian). It is equally important to point out that the currently dominant practice of pretending $\mu_{\tilde{R}} = \mu_R$ makes the strongest Bayesian assumption of all: The TRB, and hence any Δ_r ($r \geq \tilde{R}$), is exactly zero. In this sense, using a non-trivial prior for Δ_r makes less extreme assumptions than currently done in practice.

In a nutshell, a central aim of putting a prior on Δ_r to regulate the predictive power of the covariates is to identify practical ways of ordering a set of covariates to form the filtration $\{\mathcal{F}_r, r \geq 0\}$ to achieve rapid decay of $\mathrm{E}(\Delta_r^2)$ as r increases, essentially the same goal as for stepwise regression or principal component analysis. By exploring the multi-resolution formulation we hope to identify viable alternatives to common approaches such as LASSO. In general, for the multi-resolution framework to be fruitful beyond the conceptual level, many fundamental and methodological questions must be answered. The three questions below are merely antipasti to whet your appetite (for NP, or not):

(a) For what classes of models on $\{Y, X_j, j = 0, \ldots, R\}$ and priors on ordering and predictive power, can we determine practically an order $\{X_{(j)}, j \geq 0\}$ such that the resulting $\mathcal{F}_r = \sigma(X_{(j)}, j = 0, \ldots, r)$ will ensure a parsimonious representation of μ_R with quantifiably high probability?

(b) What should be our guiding principles for making a trade-off between sample size n and recorded/measured data resolution R, when we have the choice between having more data of lower quality (large n, small R) or less data of higher quality (small n, large R)?

(c) How do we determine the appropriate resolution level for hypothesis testing, considering that hypotheses testing involving higher resolution estimands typically lead to larger multiplicity? How much multiplicity can we reasonably expect our data to accommodate, and how do we quantify it?

45.3 Multi-phase inference

Most of us learned about statistical modelling in the following way. We have a data set that can be described by a random variable Y, which can be modelled by a probability function or density $\Pr(Y|\theta)$. Here θ is a model parameter, which can be of infinite dimension when we adopt a non-parametric or semi-parametric philosophy. Many of us were also taught to resist the temptation of using a model just because it is convenient, mentally, mathematically, or computationally. Instead, we were taught to learn as much as possible about the data generating process, and think critically about what makes sense substantively, scientifically, and statistically. We were then told to check and re-check the goodness-of-fit, or rather the lack of fit, of the model to our data, and to revise our model whenever our resources (time, energy, and funding) permit.

These pieces of advice are all very sound. Indeed, a hallmark of statistics as a scientific discipline is its emphasis on critical and principled thinking about the entire process from data collection to analysis to interpretation to communication of results. However, when we take our proud way of thinking (or our reputation) most seriously, we will find that we have not practiced what we have preached in a rather fundamental way.

I wish this were merely an attention-grabbing statement like the title of my article. But the reality is that when we put down a single model $\Pr(Y|\theta)$, however sophisticated or "assumption-free," we have already simplified too much. The reason is simple. In real life, especially in this age of Big Data, the data arriving at an analyst's desk or disk are almost never the original raw data, however defined. These data have been pre-processed, often in multiple phases, because someone felt that they were too dirty to be useful, or too

large to pass on, or too confidential to let the user see everything, or all of the above! Examples range from microarrays to astrophysics; see Blocker and Meng (2013).

"So what?" Some may argue that all this can be captured by our model $\Pr(Y|\theta)$, at least in theory, if we have made enough effort to learn about the entire process. Putting aside the impossibility of learning about everything in practice (Blocker and Meng, 2013), we will see that the single-model formulation is simply not rich enough to capture reality, even if we assume that every pre-processor and analyst have done everything correctly. The trouble here is that pre-processors and analysts have different goals, have access to different data resources, and make different assumptions. They typically do not and cannot communicate with each other, resulting in separate (model) assumptions that no single probabilistic model can coherently encapsulate. We need a multiplicity of models to capture a multiplicity of incompatible assumptions.

45.3.1 Multiple imputation and uncongeniality

I learned about these complications during my study of the multiple imputation (MI) method (Rubin, 1987), where the pre-processor is the imputer. The imputer's goal was to preserve as much as possible in the imputed data the joint distributional properties of the original complete data (assuming, of course, the original complete-data samples were scientifically designed so that their properties are worthy of preservation). For that purpose, the imputer should and will use anything that can help, including confidential information, as well as powerful predictive models that may not capture the correct causal relations.

In addition, because the imputed data typically will be used for many purposes, most of which cannot be anticipated at the time of imputation, the imputation model needs to include as many predictors as possible, and be as saturated as the data and resources permit; see Meng (1994) and Rubin (1996). In contrast, an analysis model, or rather an approach (e.g., given by software), often focuses on specific questions and may involve only a (small) subset of the variables used by the imputer. Consequently, the imputer's model and the user's procedure may be uncongenial to each other, meaning that no model can be compatible with both the imputer's model and the user's procedure. The technical definitions of congeniality are given in Meng (1994) and Xie and Meng (2013), which involve embedding an analyst's procedure (often of frequentist nature) into an imputation model (typically with Bayesian flavor). For the purposes of the following discussion, two models are "congenial" if their implied imputation and analysis procedures are the same. That is, they are operationally, though perhaps not theoretically, equivalent.

Ironically, the original motivation of MI (Rubin, 1987) was a separation of labor, asking those who have more knowledge and resources (e.g., the US Census Bureau) to fix/impute the missing observations, with the hope that

subsequent analysts can then apply their favorite complete-data analysis procedures to reach valid inferences. This same separation creates the issue of uncongeniality. The consequences of uncongeniality can be severe, from both theoretical and practical points of view. Perhaps the most striking example is that the very appealing variance combining rule for MI inference derived under congeniality (and another application of the aforementioned EVE law), namely,

$$\text{var}_{\text{Total}} = \text{var}_{\text{Between-imputation}} + \text{var}_{\text{Within-imputation}} \tag{45.4}$$

can lead to seriously invalid results in the presence of uncongeniality, as reported initially by Fay (1992) and Kott (1995).

Specifically, the so-called Rubin's variance combining rule is based on (45.4), where

$$\text{var}_{\text{Between-imputation}} \quad \text{and} \quad \text{var}_{\text{Within-imputation}}$$

are estimated by $(1+m^{-1})B_m$ and $\bar{U}_m$, respectively (Rubin, 1987). Here the $(1+m^{-1})$ factor accounts for the Monte Carlo error due to finite m, B_m is the sampling variance of $\hat{\theta}^{(\ell)} \equiv \hat{\theta}_A(Y_{\text{com}}^{(\ell)})$ and $\bar{U}_m$ is the sample average of $U(Y_{\text{com}}^{(\ell)}), \ell = 1, \ldots, m$, where $\hat{\theta}_A(Y_{\text{com}})$ is the analyst's complete-data estimator for θ, $U(Y_{\text{com}})$ is its associated variance (estimator), and $Y_{\text{mis}}^{(\ell)}$ are i.i.d. draws from an imputation model $P_I(Y_{\text{mis}}|Y_{\text{obs}})$. Here, for notational convenience, we assume the complete data Y_{com} can be decomposed into the missing data Y_{mis} and observed data Y_{obs}. The left-hand side of (45.4) then is meant to be an estimator, denoted by T_m, of the variance of the MI estimator of θ, that is, $\bar{\theta}_m$, the average of $\{\hat{\theta}^{(\ell)}, \ell = 1, \ldots, m\}$.

To understand the behavior of $\bar{\theta}_m$ and T_m, let us consider a relatively simple case where the missing data are missing at random (Rubin, 1976), and the imputer does not have any additional data. Yet the imputer has adopted a Bayesian model uncongenial to the analyst's complete-data likelihood function, $P_A(Y_{\text{com}}|\theta)$, even though both contain the true data-generating model as a special case. For example, the analyst may have correctly assumed that two subpopulations share the same mean, an assumption that is not in the imputation model; see Meng (1994) and Xie and Meng (2013). Furthermore, we assume the analyst's complete-data procedure is the fully efficient MLE $\hat{\theta}_A(Y_{\text{com}})$, and $U_A(Y_{\text{com}})$, say, is the usual inverse of Fisher information.

Clearly we need to take into account both the sampling variability and imputation uncertainty, and for consistency we need to take both imputation size $m \to \infty$ and data size $n \to \infty$. That is, we need to consider replications generated by the hybrid model (note $P_I(Y_{\text{mis}}|Y_{\text{obs}})$ is free of θ):

$$P_H(Y_{\text{mis}}, Y_{\text{obs}}|\theta) = P_I(Y_{\text{mis}}|Y_{\text{obs}})P_A(Y_{\text{obs}}|\theta), \tag{45.5}$$

where $P_A(Y_{\text{obs}}|\theta)$ is derived from the analyst's complete-data model $P_A(Y_{\text{com}}|\theta)$.

To illustrate the complication caused by uncongeniality, let us assume $m = \infty$ to eliminate the distraction of Monte Carlo error due to finite m. Writing

$$\bar{\theta}_\infty - \theta = \{\bar{\theta}_\infty - \hat{\theta}_A(Y_{\rm com})\} + \{\hat{\theta}_A(Y_{\rm com}) - \theta\},$$

we have

$$\begin{aligned} {\rm var}_H(\bar{\theta}_\infty) &= {\rm var}_H\{\bar{\theta}_\infty - \hat{\theta}_A(Y_{\rm com})\} + {\rm var}_H\{\hat{\theta}_A(Y_{\rm com})\} \\ &\quad +2\,{\rm cov}_H\{\bar{\theta}_\infty - \hat{\theta}_A(Y_{\rm com}), \hat{\theta}_A(Y_{\rm com})\}, \end{aligned} \tag{45.6}$$

where all the expectations are with respect to the hybrid model defined in (45.5). Since we assume both the imputer's model and the analyst's model are valid, it is not too hard to see intuitively — and to prove under regularity conditions, as in Xie and Meng (2013) — that the first term and second term on the right-hand side of (45.6) are still estimated consistently by B_m and $\bar{U}_m$, respectively. However, the trouble is that the cross term as given in (45.6) is left out by (45.4), so unless this term is asymptotically negligible, Rubin's variance estimator of ${\rm var}_H(\bar{\theta}_\infty)$ via (45.4) cannot be consistent, an observation first made by Kott (1995).

Under congeniality, this term is indeed negligible. This is because, under our current setting, $\bar{\theta}_\infty$ is asymptotically (as $n \to \infty$) the same as the analyst's MLE based on the observed data $Y_{\rm obs}$; we denote it, with an abuse of notation, by $\hat{\theta}_A(Y_{\rm obs})$. But $\hat{\theta}_A(Y_{\rm obs}) - \hat{\theta}_A(Y_{\rm com})$ and $\hat{\theta}_A(Y_{\rm com})$ must be asymptotically orthogonal (i.e., uncorrelated) under P_A, which in turn is asymptotically the same as P_H due to congeniality (under the usual regularity conditions that guarantee the equivalence of frequentist and Bayesian asymptotics). Otherwise there must exist a linear combination of $\hat{\theta}_A(Y_{\rm obs}) - \hat{\theta}_A(Y_{\rm com})$ and $\hat{\theta}_A(Y_{\rm com})$ — and hence of $\hat{\theta}_A(Y_{\rm obs})$ and $\hat{\theta}_A(Y_{\rm com})$ — that is asymptotically more efficient than $\hat{\theta}_A(Y_{\rm com})$, contradicting the fact that $\hat{\theta}_A(Y_{\rm com})$ is the full MLE under $P_A(Y_{\rm com}|\theta)$.

When uncongeniality arises, it becomes entirely possible that there exists a linear combination of $\bar{\theta}_\infty - \hat{\theta}_A(Y_{\rm com})$ and $\hat{\theta}_A(Y_{\rm com})$ that is more efficient than $\theta_A(Y_{\rm com})$ at least under the actual data generating model. This is because $\bar{\theta}_\infty$ may inherit, through the imputed data, additional (valid) information that is not available to the analyst, and hence is not captured by $P_A(Y_{\rm com}|\theta)$. Consequently, the cross-term in (45.6) is not asymptotically negligible, making (45.4) an inconsistent variance estimator; see Fay (1992), Meng (1994) and Kott (1995).

The above discussion also hints at an issue that makes the multi-phase inference formulation both fruitful and intricate, because it indicates that consistency can be preserved when the imputer's model does not bring in additional (correct) information. This is a much weaker requirement than congeniality, because it is satisfied, for example, when the analyst's model is nested within (i.e., less saturated than) the imputer's model. Indeed, in Xie and Meng (2013) we established precisely this fact, under regularity conditions. However, when we assume that the imputer model is nested within the analyst's model, we

can prove only that (45.4) has a positive bias. But even this weaker result requires an additional assumption — for multivariate θ — that the loss of information is the same for all components of θ. This additional requirement for multivariate θ was both unexpected and troublesome, because in practice there is little reason to expect that the loss of information will be the same for different parameters.

All these complications vividly demonstrate both the need for and challenges of the multi-phase inference framework. By multi-phase, our motivation is not merely that there are multiple parties involved, but more critically that the phases are sequential in nature. Each phase takes the output of its immediate previous phase as the input, but with little knowledge of how other phases operate. This lack of mutual knowledge reality leads to uncongeniality, which makes any single-model framework inadequate for reasons stated before.

45.3.2 Data pre-processing, curation and provenance

Taking this multi-phase perspective but going beyond the MI setting, we (Blocker and Meng, 2013) recently explored the steps needed for building a theoretical foundation for pre-processing in general, with motivating applications from microarrays and astrophysics. We started with a simple but realistic two-phase setup, where for the pre-processor phase, the input is Y and the output is $T(Y)$, which becomes the input of the analysis phase. The pre-process is done under an "observation model" $P_Y(Y|X,\xi)$, where X represents the ideal data we do not have (e.g., true expression level for each gene), because we observe only a noisy version of it, Y (e.g., observed probe-level intensities), and where ξ is the model parameter characterizing how Y is related to X, including how noises were introduced into the observation process (e.g., background contamination). The downstream analyst has a "scientific model" $P_X(X|\theta)$, where θ is the scientific estimand of interest (e.g., capturing the organism's patterns of gene expression). To the analyst, both X and Y are missing, because only $T(Y)$ is made available to the analyst. For example, $T(Y)$ could be background corrected, normalized, or aggregated Y. The analyst's task is then to infer θ based on $T(Y)$ only.

Given such a setup, an obvious question is what $T(Y)$ should the pre-processor produce/keep in order to ensure that the analyst's inference of θ will be as sharp as possible? If we ignore practical constraints, the answer seems to be rather trivial: Choose $T(Y)$ to be a (minimal) sufficient statistic for

$$P_Y(y|\theta,\xi) = \int P_Y(y|x;\xi)P_X(x|\theta)\mu(\mathrm{d}x). \tag{45.7}$$

But this does not address the real problem at all. There are thorny issues of dealing with the nuisance (to the analyst) parameter ξ, as well as the issue of computational feasibility and cost. But most critically, because of the separation of the phases, the scientific model $P_X(X|\theta)$ and hence the marginal

model $P_Y(Y|\theta,\xi)$ of (45.7) is typically unknown to the pre-processor. At the very best, the pre-processor may have a working model $\tilde{P}_X(X|\eta)$, where η may not live even on the same space as θ. Consequently, the pre-processor may produce $T(Y)$ as a (minimal) sufficient statistic with respect to

$$\tilde{P}_Y(y|\eta,\xi) = \int P_Y(y|x;\xi)\tilde{P}_X(x|\eta)\mu(\mathrm{d}x). \tag{45.8}$$

A natural question then is what are sufficient and necessary conditions on the pre-processor's working model such that a $T(Y)$ (minimally) sufficient for (45.8) will also be (minimally) sufficient for (45.7). Or to use computer science jargon, when is $T(Y)$ a lossless compression (in terms of statistical efficiency)?

Evidently, we do not need the multi-phase framework to obtain trivial and useless answers such as setting $T(Y) = Y$ (which will be sufficient for any model of Y only) or requiring the working model to be the same as the scientific model (which tells us nothing new). The multi-phase framework allows us to formulate and obtain theoretically insightful and practically relevant results that are unavailable in the single-phase framework. For example, in Blocker and Meng (2013), we obtained a non-trivial sufficient condition as well as a necessary condition (but they are not the same) for preserving sufficiency under a more general setting involving multiple (parallel) pre-processors during the pre-process phase. The sufficient condition is in the same spirit as the condition for consistency of Rubin's variance rule under uncongeniality. That is, in essence, sufficiency under (45.8) implies sufficiency under (45.7) when the working model is more saturated than the scientific model. This is rather intuitive from a multi-phase perspective, because the fewer assumptions we make in earlier phases, the more flexibility the later phases inherit, and consequently, the better the chances these procedures preserve information or desirable properties.

There is, however, no free lunch. The more saturated our model is, the less compression it achieves by statistical sufficiency. Therefore, in order to make our results as practically relevant as possible, we must find ways to incorporate computational efficiency into our formulation. However, establishing a general theory for balancing statistical and computational efficiency is an extremely challenging problem. The central difficulty is well known: Statistical efficiency is an inherent property of a procedure, but the computational efficiency can vary tremendously across computational architectures and over time.

For necessary conditions, the challenge is of a different kind. Preserving sufficiency is a much weaker requirement than preserving a model, even for minimal sufficiency. For example, $\mathcal{N}(\mu,1)$ and Poisson(λ) do not share even the same state space. However, the sample mean is a minimal sufficient statistic for both models. Therefore, a pre-processing model could be seriously flawed yet still lead to the best possible pre-processing (this could be viewed as a case of action consistency; see Section 45.5). This type of possibility makes building a multi-phase inference theory both intellectually demanding and intriguing.

In general, "What to keep?" or "Who will share what, with whom, when, and why?" are key questions for the communities in information and computer sciences, particularly in the areas of data curation and data provenance; see Borgman (2010) and Edwards et al. (2011). Data/digital curation, as defined by the US National Academies, is "the active management and enhancement of digital information assets for current and future use," and data provence is "a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing"; see Moreau et al. (2013). Whereas these fields are clearly critical for preserving data quality and understanding the data collection process for statistical modelling, currently there is little dialogue between these communities and statisticians despite shared interests. For statisticians to make meaningful contributions, we must go beyond the single-phase/single-model paradigm because the fundamental problems these fields address involve, by default, multiple parties, who do not necessarily (or may not even be allowed to) share information, and yet they are expected to deliver scientifically useful data and digital information.

I believe the multi-phase inference framework will provide at least a relevant formulation to enter the conversation with researchers in these areas. Of course, there is a tremendous amount of foundation building to be done, even just to sort out which results in the single-phase framework are directly transferable and which are not. The three questions below again are just an appetizer:

(a) What are practically relevant theoretical criteria for judging the quality of pre-processing, without knowing how many types of analyses ultimately will be performed on the pre-processed data?

(b) What are key considerations and methods for formulating generally uncongeniality for multi-phase inference, for quantifying the degrees of uncongeniality, and for setting up a threshold for a tolerable degree?

(c) How do we quantify trade-offs between efficiencies that are designed for measuring different aspects of the multi-phase process, such as computational efficiency for pre-processing and statistical efficiency for analysis?

45.4 Multi-source inference

As students of statistics, we are all taught that a scientific way of collecting data from a population is to take a probabilistic sample. However, this was not the case a century ago. It took about half a century since its formal introduction in 1895 by Anders Nicolai Kiær (1838–1919), the founder of Statistics Norway, before probabilistic sampling became widely understood

and accepted (see Bethlehem, 2009). Most of us now can explain the idea intuitively by analogizing it with common practices such as that only a tiny amount of blood is needed for any medical test (a fact for which we are all grateful). But it was difficult then for many — and even now for some — to believe that much can be learned about a population by studying only, say, a 5% random sample. Even harder was the idea that a 5% random sample is better than a 5% "quota sample," i.e., a sample purposefully chosen to mimic the population. (Very recently a politician dismissed an election pool as "non-scientific" because "it is random.")

Over the century, statisticians, social scientists, and others have amply demonstrated theoretically and empirically that (say) a 5% probabilistic/random sample is better than any 5% non-random samples in many measurable ways, e.g., bias, MSE, confidence coverage, predictive power, etc. However, we have not studied questions such as "Is an 80% non-random sample 'better' than a 5% random sample in measurable terms? 90%? 95%? 99%?"

This question was raised during a fascinating presentation by Dr. Jeremy Wu, then (in 2009) the Director of LED (Local Employment Dynamic), a pioneering program at the US Census Bureau. LED employed synthetic data to create an OnTheMap application that permits users to zoom into any local region in the US for various employee-employer paired information without violating the confidentiality of individuals or business entities. The synthetic data created for LED used more than 20 data sources in the LEHD (Longitudinal Employer-Household Dynamics) system. These sources vary from survey data such as a monthly survey of 60,000 households, which represent only .05% of US households, to administrative records such as unemployment insurance wage records, which cover more than 90% of the US workforce, to census data such as the quarterly census of earnings and wages, which includes about 98% of US jobs (Wu, 2012 and personal communication from Wu).

The administrative records such as those in LEHD are not collected for the purpose of statistical inference, but rather because of legal requirements, business practice, political considerations, etc. They tend to cover a large percentage of the population, and therefore they must contain useful information for inference. At the same time, they suffer from the worst kind of selection biases because they rely on self-reporting, convenient recording, and all sorts of other "sins of data collection" that we tell everyone to avoid.

But statisticians cannot avoid dealing with such complex combined data sets, because they are playing an increasingly vital role for official statistical systems and beyond. For example, the shared vision from a 2012 summit meeting, between the government statistical agencies from Australia, Canada, New Zealand, the United Kingdom, and the US, includes

> "Blending together multiple available data sources (administrative and other records) with traditional surveys and censuses (using paper, internet, telephone, face-to-face interviewing) to create high quality, timely statistics that tell a coherent story of economic, social and en-

vironmental progress must become a major focus of central government statistical agencies." (Groves, February 2, 2012)

Multi-source inference therefore refers to situations where we need to draw inference by using data coming from different sources and some (but not all) of which were not collected for inference purposes. It is thus broader and more challenging than multi-frame inference, where multiple data sets are collected for inference purposes but with different survey frames; see Lohr and Rao (2006). Most of us would agree that the very foundation of statistical inference is built upon having a representative sample; even in notoriously difficult observational studies, we still try hard to create pseudo "representative" samples to reduce the impact of confounding variables. But the availability of a very large subpopulation, however biased, poses new opportunities as well as challenges.

45.4.1 Large absolute size or large relative size?

Let us consider a case where we have an administrative record covering f_a percent of the population, and a simple random sample (SRS) from the same population which only covers f_s percent, where $f_s << f_a$. Ideally, we want to combine the maximal amount of information from both of them to reach our inferential conclusions. But combining them effectively will depend critically on the relative information content in them, both in terms of how to weight them (directly or implied) and how to balance the gain in information with the increased analysis cost. Indeed, if the larger administrative dataset is found to be too biased relative to the cost of processing it, we may decide to ignore it. Wu's question therefore is a good starting point because it directly asks how the relative information changes as their relative sizes change: How large should f_a/f_s be before an estimator from the administrative record dominates the corresponding one from the SRS, say in terms of MSE?

As an initial investigation, let us denote our finite population by $\{x_1, \ldots, x_N\}$. For the administrative record, we let $R_i = 1$ whenever x_i is recorded and zero otherwise; and for SRS, we let $I_i = 1$ if x_i is sampled, and zero otherwise, $i = 1, \ldots, N$. Here we assume $n_a = \sum_{i=1}^{N} R_i >> n_s = \sum_{i=1}^{N} I_i$, and both are considered fixed in the calculations below. Our key interest here is to compare the MSEs of two estimators of the finite-sample population mean $\bar{X}_N$, namely,

$$\bar{x}_a = \frac{1}{n_a} \sum_{i=1}^{N} x_i R_i \quad \text{and} \quad \bar{x}_s = \frac{1}{n_s} \sum_{i=1}^{N} x_i I_i.$$

Recall for finite-population calculations, all x_i's are fixed, and all the randomness comes from the response/recording indicator R_i for $\bar{x}_a$ and the sampling indicator I_i for $\bar{x}_s$. Although the administrative record has no probabilistic mechanism imposed by the data collector, it is a common strategy to model the responding (or recording or reporting) behavior via a probabilistic model.

Here let us assume that a probit regression model is adequate to capture the responding behavior, which depends on only the individual's x value. That is, we can express $R_i = 1_{\{Z_i \le \alpha+\beta x_i\}}$, where Z_i's are i.i.d samples from $\mathcal{N}(0,1)$. We could imagine Z_i being, e.g., the ith individual's latent "refusal tendency," and when it is lower than a threshold that is linear in x_i, the individual responds. The intercept α allows us to model the overall percentage of respondents, with larger α implying more respondents. The slope β models the strength of the self-selecting mechanism. In other words, as long as $\beta \neq 0$, we have a non-ignorable missing-data mechanism (Rubin, 1976).

Given that $\bar{x}_s$ is unbiased, its MSE is the same as its variance (Cochran, 2007), viz.

$$\mathrm{var}(\bar{x}_s) = \frac{1-f_s}{n_s} S_N^2(x), \quad \text{where } S_N^2(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x}_N)^2. \tag{45.9}$$

The MSE of $\bar{x}_a$ is more complicated, mostly because R_i depends on x_i. But under our assumption that N is very large and $f_a = n_a/N$ stays (far) away from zero, the MSE is completely dominated by the squared bias term of $\bar{x}_a$, which itself is well approximated by, again because N (and hence n_a) is very large,

$$\mathrm{Bias}^2(\bar{x}_a) = \left\{ \frac{\sum_{i=1}^{N}(x_i - \bar{x}_N)p(x_i)}{\sum_{i=1}^{N} p(x_i)} \right\}^2, \tag{45.10}$$

where $p(x_i) = \mathrm{E}(R_i|x_i) = \Phi(\alpha + \beta x_i)$, and $\Phi(Z)$ is the CDF for $\mathcal{N}(0,1)$.

To get a sense of how this bias depends on f_a, let us assume that the finite population $\{x_1, \ldots, x_N\}$ itself can be viewed as a size N SRS from a super population $X \sim \mathcal{N}(\mu, \sigma^2)$. By the law of large number, the bias term in (45.10) is essentially the same as (again because N is very large)

$$\frac{\mathrm{cov}\{X, p(X)\}}{\mathrm{E}\{p(X)\}} = \frac{\sigma \mathrm{E}\{Z\Phi(\tilde{\alpha}+\tilde{\beta}Z)\}}{\mathrm{E}\{\Phi(\tilde{\alpha}+\tilde{\beta}Z)\}} = \frac{\sigma\tilde{\beta}}{\sqrt{1+\tilde{\beta}^2}} \frac{\phi\left(\dfrac{\tilde{\alpha}}{\sqrt{1+\tilde{\beta}^2}}\right)}{\Phi\left(\dfrac{\tilde{\alpha}}{\sqrt{1+\tilde{\beta}^2}}\right)}, \tag{45.11}$$

where $\tilde{\alpha} = \alpha + \beta\mu, \tilde{\beta} = \sigma\beta$, $Z \sim \mathcal{N}(0,1)$, and $\phi(z)$ is its density function. (Integration by parts and properties of normals are used for arriving at (45.11).)

An insight is provided by (45.11) when we note $\Phi\{\tilde{\alpha}/(1+\tilde{\beta}^2)^{1/2}\}$ is well estimated by f_a because N is large, and hence $\tilde{\alpha}/(1+\tilde{\beta}^2)^{1/2} \approx \Phi^{-1}(f_a) = z_{f_a}$, where z_q is the qth quantile of $\mathcal{N}(0,1)$. Consequently, we have from (45.11),

$$\frac{\mathrm{MSE}(\bar{x}_a)}{\sigma^2} \approx \frac{\mathrm{Bias}^2(\bar{x}_a)}{\sigma^2} = \frac{\tilde{\beta}^2}{1+\tilde{\beta}^2}\frac{\phi^2(z_{f_a})}{f_a^2} = \frac{\tilde{\beta}^2}{1+\tilde{\beta}^2}\frac{e^{-z_{f_a}^2}}{2\pi f_a^2}, \tag{45.12}$$

which will be compared to (45.9) after replacing $S_N^2(X)$ by σ^2. That is,

$$\frac{\text{MSE}(\bar{x}_s)}{\sigma^2} = \frac{1}{n_s} - \frac{1}{N} \approx \frac{1}{n_s}, \tag{45.13}$$

where $1/N$ is ignored for the same reason that $\text{var}(\bar{x}_a) = O(N^{-1})$ is ignored.

It is worthy to point out that the seemingly mismatched units in comparing (45.12), which uses relative size f_a, with (45.13), which uses the absolute size n_s, reflects the different natures of non-sampling and sampling errors. The former can be made arbitrarily small only when the relative size f_a is made arbitrarily large, that is $f_a \to 1$; just making the absolute size n_a large will not do the trick. In contrast, as is well known, we can make (45.13) arbitrarily small by making the absolute size n_s arbitrarily large even if $f_s \to 0$ when $N \to \infty$. Indeed, for most public-use data sets, f_s is practically zero. For example, with respect to the US population, an $f_s = .01\%$ would still render n_s more than 30,000, large enough for controlling sampling errors for many practical purposes. Indeed, (45.13) will be no greater than .000033. In contrast, if we were to use an administrative record of the same size, that is, if $f_a = .01\%$, then (45.12) will be greater than 3.13, almost 100,000 times (45.13), if $\tilde{\beta} = .5$.

However, if $f_a = 95\%$, $z_{f_a} = 1.645$, (45.12) will be .00236, for the same $\tilde{\beta} = .5$. This implies that as long as n_s does not exceed about 420, the estimator from the biased sample will have a smaller MSE (assuming, of course, $N >> 420$). The threshold value for n_s will drop to about 105 if we increase $\tilde{\beta}$ to 2, but will increase substantially to about 8,570 if we drop $\tilde{\beta}$ to .1. We must be mindful, however, that these comparisons assume the SRS and more generally the survey data have been collected perfectly, which will not be the case in reality because of both non-responses and response biases; see Liu et al. (2013). Hence in reality it would take a smaller f_a to dominate the probabilistic sample with f_s sampling fraction, precisely because the latter has been contaminated by non-probabilistic selection errors as well. Nevertheless, a key message here is that, as far as statistical inference goes, what makes a "Big Data" set big is typically not its absolute size, but its relative size to its population.

45.4.2 Data defect index

The sensitivity of our comparisons above to $\tilde{\beta}$ is expected because it governs the self-reporting mechanism. In general, whereas closed-form expressions such as (45.12) are hard to come by, the general expression in (45.10) leads to

$$\frac{\text{Bias}^2(\bar{x}_a)}{S_N^2(x)} = \rho_N^2(x,p) \left\{ \frac{S_N(p)}{\bar{p}_N} \right\}^2 \left\{ \frac{N-1}{N} \right\}^2 < \rho_N^2(x,p) \frac{1-\bar{p}_N}{\bar{p}_N}, \tag{45.14}$$

where $\bar{p}_N$ is the mean of p_i, $\rho_N(x,p)$ is the correlation between x_i and p_i, and the term inside the first set of brackets is the coefficient of variation of p_i, all of which are with respect to the finite population, that is, the uniform

distribution over the index space $\{i = 1, \ldots, N\}$. This explains the notation $\rho_N(x, p)$, in contrast to $\rho(X, p(X))$, which is with respect to X from the super population.

The (middle) re-expression of the bias given in (45.14) in terms of the correlation between sampling variable x and sampling/response probability p is a standard strategy in the survey literature; see Hartley and Ross (1954) and Meng (1993). Although mathematically trivial, it provides a greater statistical insight, that is, the sample mean from an arbitrary sample is an unbiased estimator for the target population mean if and only if the sampling variable x and the data collection mechanism $p(x)$ are uncorrelated. In this sense we can view $\rho_N(x, p)$ as a "defect index" for estimation (using sample mean) due to the defect in data collection/recording. This result says that we can reduce estimation bias of the sample mean for non-equal probability samples or even non-probability samples as long as we can reduce the magnitude of the correlation between x and $p(x)$. This possibility provides an entryway into dealing with a large but biased sample, and exploiting it may require less knowledge about $p(x)$ than required for other bias reduction techniques such as (inverse probability) weighting, as in the Horvitz-Thompson estimator.

The (right-most) inequality in (45.14) is due to the fact that for any random variable satisfying $U \in [0, 1]$, $\text{var}(U) \leq \text{E}(U)\{1 - \text{E}(U)\}$. This bound allows us to control the bias using only the proportion $\bar{p}_N$, which is well estimated by the observed sample fraction f_a. It says that we can also control the bias by letting f_a approach one. In the traditional probabilistic sampling context, this observation would only induce a "duhhh" response, but in the context of multi-source inference it is actually a key reason why an administrative record can be very useful despite being a non-probabilistic sample.

Cautions are much needed however, because (45.14) also indicates that it is not easy at all to use a large f_a to control the bias (and hence MSE). By comparing (45.13) and the bound in (45.14) we will need (as a sufficient condition)

$$f_a > \frac{n_s \rho_N^2(x, p)}{1 + n_s \rho_N^2(x, p)}$$

in order to guarantee $\text{MSE}(\bar{x}_a) < \text{MSE}(\bar{x}_s)$. For example, even if $n_s = 100$, we would need over 96% of the population if $\rho_N = .5$. This reconfirms the power of probabilistic sampling and reminds us of the danger in blindly trusting that "Big Data" must give us better answers. On the other hand, if $\rho_N = .1$, then we will need only 50% of the population to beat a SRS with $n_s = 100$. If $n_s = 100$ seems too small in practice, the same $\rho_N = .1$ also implies that a 96% subpopulation will beat a SRS as large as $n_s = \rho_N^{-2}\{f_a/(1 - f_a)\} = 2400$, which is no longer a practically irrelevant sample size.

Of course all these calculations depend critically on knowing the value of ρ_N, which cannot be estimated from the biased sample itself. However, recall for multi-source inference we will also have at least a (small) probabilistic sample. The availability of both small random sample(s) and large non-random

sample(s) opens up many possibilities. The following (non-random) sample of questions touch on this and other issues for multi-source inference:

(a) Given partial knowledge of the recording/response mechanism for a (large) biased sample, what is the optimal way to create an intentionally biased sub-sampling scheme to counter-balance the original bias so the resulting sub-sample is guaranteed to be less biased than the original biased sample in terms of the sample mean, or other estimators, or predictive power?

(b) What should be the key considerations when combining small random samples with large non-random samples, and what are the sensible "corner-cutting" guidelines when facing resource constraints? How can the combined data help to estimate $\rho_N(x, p)$? In what ways can such estimators aid multi-source inference?

(c) What are theoretically sound and practically useful defect indices for prediction, hypothesis testing, model checking, clustering, classification, etc., as counterparts to the defect index for estimation, $\rho_N(x, p)$? What are their roles in determining information bounds for multi-source inference? What are the relevant information measures for multi-source inference?

45.5 The ultimate prize or price

Although we have discussed the trio of inference problems separately, many real-life problems involve all of them. For example, the aforementioned On-TheMap application has many resolution levels (because of arbitrary zoom-in), many sources of data (more than 20 sources), and many phases of pre-process (even God would have trouble keeping track of all the processing that these twenty some survey, census, and administrative data sets have endured!), including the entire process of producing the synthetic data themselves. Personalized medicine is another class of problems where one typically encounters all three types of complications. Besides the obvious resolution issue, typically the data need to go through pre-processing in order to protect the confidentiality of individual patients (beyond just removing the patient's name). Yet individual level information is most useful. To increase the information content, we often supplement clinical trial data with observational data, for example, on side effects when the medications were used for another disease.

To bring the message home, it is a useful exercise to imagine ourselves in a situation where our statistical analysis would actually be used to decide the best treatment for a serious disease for a loved one or even for ourselves. Such a "personalized situation" emphasizes that it is my interest/life at stake, which should encourage us to think more critically and creatively, not just to publish another paper or receive another prize. Rather, it is about getting to

the bottom of what we do as statisticians — to transform whatever empirical observations we have into the best possible quantitative evidence for scientific understanding and decision making, and more generally, to advance science, society, and civilization. That is our ultimate prize.

However, when we inappropriately formulate our inference problems for mental, mathematical, or computational convenience, the chances are that someone or, in the worst case, our entire society will pay the ultimate price. We statisticians are quick to seize upon the 2008 world-wide financial crisis as an ultimate example in demonstrating how a lack of understanding and proper accounting for uncertainties and correlations leads to catastrophe. Whereas this is an extreme case, it is unfortunately not an unnecessary worry that if we continue to teach our students to think only in a single-resolution, single-phase, single-source framework, then there is only a single outcome: They will not be at the forefront of quantitative inference. When the world is full of problems with complexities far exceeding what can be captured by our theoretical framework, our reputation for critical thinking about the entirety of the inference process, from data collection to scientific decision, cannot stand.

The "personalized situation" also highlights another aspect that our current teaching does not emphasize enough. If you really had to face the unfortunate I-need-treatment-now scenario, I am sure your mind would not be (merely) on whether the methods you used are unbiased or consistent. Rather, the type of questions you may/should be concerned with are (1) "Would I reach a different conclusion if I use another analysis method?" or (2) "Have I really done the best given my data and resource constraints?" or (3) "Would my conclusion change if I were given all the original data?"

Questions (1) and (2) remind us to put more emphasis on relative optimality. Whereas it is impossible to understand all biases or inconsistencies in messy and complex data, knowledge which is needed to decide on the optimal method, we still can and should compare methods relative to each other, as well as relative to the resources available (e.g., time, energy, funding). Equally important, all three questions highlight the need to study much more qualitative consistency or action consistency than quantitative consistency (e.g., the numerical value of our estimator reaching the exact truth in the limit). Our methods, data sets, and numerical results can all be rather different (e.g., a p-value of .2 versus .8), yet their resulting decisions and actions can still be identical because typically there are only two (yes and no) or at most a handful of choices.

It is this "low resolution" of our action space in real life which provides flexibility for us to accept quantitative inconsistency caused by defects such as resolution discrepancy, uncongeniality or selection bias, yet still reach scientifically useful inference. It permits us to move beyond single-phase, single-source, or single resolution frameworks, but still be able to obtain theoretically elegant and practically relevant results in the same spirit as those NP-worthy findings in many other fields. I therefore very much hope you will join me for

this intellectually exciting and practically rewarding research journey, unless, of course, you are completely devoted to fundraising to establish an NP in statistics.

Acknowledgements

The material on multi-resolution inference benefitted greatly from critical comments by Alex Blocker and Keli Liu, both of whom also provided many insightful comments throughout, as did David Jones. The joint work with Alex Blocker and Xianchao Xie (cited in the reference list) shaped the formulation of the multi-phase inference, which was greatly encouraged by Christine Borgman, who also taught me, together with Alyssa Goodman, Paul Groth, and Margaret Hedstrom, data curation and data provenance. Dr. Jeremy Wu inspired and encouraged me to formulate the multi-source inference, and provided extremely helpful information and insights regarding the LED/LEHD program. Keli Liu also provided invaluable editing and proofreading, as did Steven Finch. "Good stuff!" coming from my academic twin brother Andrew Gelman was all the encouragement I needed to squeeze out every possible minute between continental breakfasts and salmon/chicken dinners. I give them 100% thanks, but 0% liability for any naïveté, wishful thinking, and sign of lack of sleep — this has been the most stressful paper I have ever written. I also thank the NSF for partial financial support, and the co-Editors, especially Xihong Lin and Geert Molenberghs, for help and extraordinary patience.

References

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Bethlehem, J. (2009). *The Rise of Survey Sampling*. CBS Discussion Paper No. 9015.

Blitzstein, J. and Meng, X.-L. (2010). Nano-project qualifying exam process: An intensified dialogue between students and faculty. *The American Statistician*, 64:282–290.

Blocker, A.W. and Meng, X.-L. (2013). The potential and perils of preprocessing: Building new foundations. *Bernoulli*, 19:1176–1211.

Borgman, C.L. (2010). Research data: Who will share what, with whom, when, and why? *China-North America Library Conference, Beijing*, People's Republic of China.

Bouman, P., Dukic, V. and Meng, X.-L. (2005). A Bayesian multiresolution hazard model with application to an AIDS reporting delay study. *Statistica Sinica*, 15:325–357.

Bouman, P., Meng, X.-L., Dignam, J., and Dukić, V. (2007). A multiresolution hazard model for multicenter survival studies: Application to tamoxifen treatment in early stage breast cancer. *Journal of the American Statistical Association*, 102:1145–1157.

Cochran, W.G. (2007). *Sampling Techniques*. Wiley, New York.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM.

Donoho, D.L. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100:2197–2202.

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: Asymptopia? (with discussion). *Journal of the Royal Statistical Society, Series B*, 57:301–369.

Edwards, P.N., Mayernik, M.S., Batcheller, A.L., Bowker, G.C., and Borgman, C.L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41:667–690.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section*, American Statistical Association, Washington, DC, pp. 227–232.

Garey, M. (1974). Optimal binary search trees with restricted maximal depth. *SIAM Journal on Computing*, 3:101–110.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis*. Chapman & Hall, London.

Gelman, A. and Little, T.C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23:127–35.

George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373.

Groves, R.M. (February 2, 2012). National statistical offices: Independent, identical, simultaneous actions thousands of miles apart US Census Bureau Director's Blog, `http://blogs.census.gov/directorsblog/`.

Hartley, H. and Ross, A. (1954). Unbiased ratio estimators. *Nature*, 174:270–271.

Hirakawa, K. and Meng, X.-L. (2006). An empirical Bayes EM-wavelet unification for simultaneous denoising, interpolation, and/or demosaicing. In *Image Processing, 2006 IEEE International Conference on*. IEEE, pp. 1453–1456.

Knuth, D. (1997). *The Art of Computer Programming*, Vol 1. *Fundamental Algorithms*, 3rd edition. Addison-Wesley, Reading, MA.

Kott, P.S. (1995). A paradox of multiple imputation. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Washington, DC, pp. 380–383.

Lax, J.R. and Phillips, J.H. (2009). How should we estimate public opinion in the states? *American Journal of Political Science*, 53:107–121.

Lee, T.C. and Meng, X.-L. (2005). A self-consistent wavelet method for denoising images with missing pixels. In *Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:41–44.

Lin, D., Foster, D.P., and Ungar, L.H. (2010). *A Risk Ratio Comparison of ℓ_0 and ℓ_1 Penalized Regressions*. Technical Report, University of Pennsylvania, Philadelphia, PA.

Liu, J., Meng, X.-L., Chen, C.-N. and Alegrita, M. (2013). Statistics can lie but can also correct for lies: Reducing response bias in NLAAS via Bayesian imputation. *Statistics and Its Interface*, 6:387–398.

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

Meng, X.-L. (1993). On the absolute bias ratio of ratio estimators. *Statistics & Probability Letters*, 18:345–348.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 9:538–558.

Meng, X.-L. (2009a). Automated bias-variance trade-off: Intuitive inadmissibility or inadmissible intuition? In *Frontiers of Statistical Decision Making and Bayesian Analysis* (M.H. Chen, D.K. Dey, P. Mueller, D. Sun, and K. Ye, Eds.). Springer, New York, pp. 95–112.

Meng, X.-L. (2009b). Desired and feared — What do we do now and over the next 50 years? *The American Statistician*, 63:202–210.

Meyer, Y. (1993). Wavelets-algorithms and applications. *Wavelets-Algorithms and Applications*, 1:142.

Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032.

Moreau, L., Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., and Tilmes, C., Eds (2013). *PROV-DM: The PROV Data Model*. Technical Report, World Wide Web Consortium.

Nason, G.P. (2002). Choice of wavelet smoothness, primary resolution and threshold in wavelet shrinkage. *Statistics and Computing*, 12:219–227.

Poggio, T. and Girosi, F. (1998). A sparse representation for function approximation. *Neural Computation*, 10:1445–1454.

Rao, J.N.K. (2005). *Small Area Estimation*. Wiley, New York.

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

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489.

Rubin, D.B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100:322–331.

Wu, J. (2012). 21st century statistical systems. *Blog: NotRandomThought*, August 1, 2012. Available at `http://jeremyswu.blogspot.com/`.

Xie, X. and Meng, X.-L. (2013). Dissecting multiple imputation from a multiphase inference perspective: What happens when there are three uncongenial models involved? *The Annals of Statistics*, under review.

Yang, J., Peng, Y., Xu, W., and Dai, Q. (2009). Ways to sparse representation: An overview. *Science in China Series F: Information Sciences*, 52:695–703.