The New England Journal of Statistics in Data Science Volume 1, 31–34 (2023)
DOI: https://doi.org/10.51387/23-NEJSDS6E

Comments on Xiao-Li Meng's Double Your Variance, Dirtify Your Bayes, Devour Your Pufferfish, and Draw Your Kidstogram☆

Dennis K.J. LIN

Professor Xiao-Li Meng is to be congratulated for this thoughtful and potentially highly impactful article towards improving the reproducibility and reliability of scientific studies. This paper is very well written and interesting to read. This article proposes four "radical" changes in the practice of statistics with the overall goal of increasing public trust in statistics and, by extension, scientific research overall. There is a great emphasis on the idea that statistical analyses should come with a "quality guarantee," e.g., if a study produces a 95% confidence interval, then we want to be sure the error is at most 5% (ideally less than 5%) and this result should be trusted, reproducible for future studies, and more easily understood by the public. In short, it emphasizes "Don't sell whatever you refuse to buy."

To this end, the author discusses the following four ideas:

1. We should double the variance when performing inference. The author argues that this will provide more robustness against incorrect assumptions and model misspecification. (Section 2)
2. We should "cut corners" (use approximations and simplifying assumptions) in a way that addresses practical constraints while also putting clear bounds on the approximation. (Section 3)
3. We should not sell statistical methods that we would not use for our own analysis. This requires some form of "quality introspection" before finishing an analysis. (Section 4)
4. We should teach children about distributions and the concept of variation at an early age, so that they are better equipped for a world of uncertainty. (Section 5)

These ideas are very well communicated.

My main concerns:

1. "Validity of Inference" vs. "Quality of Inference": the quality of inference should be characterized by both validity and efficiency. Inference of high quality should be made in principled ways, at least in discussions in the academic world.
2. Some ideas are still at the stage of "proposal" and hard to be implemented in practice. For example, doubling variance leads to a larger p-value, but the enlarged p-value is difficult to interpret under the existing probabilistic framework.
3. The sample size (more precisely, the effective sample size) matters for the reliability of the p-value, as the populations underlying our studies, especially for social and medical studies, are often heterogeneous. Studies with a small effective sample size can likely lead to a biased inference for a heterogeneous population. We have many examples for population heterogeneity. The most famous one is perhaps cancer, where heterogeneity is not limited to differences between different patients but also within a single patient (Allison and Sledge, 2014) [1]. The observation of the heterogeneity of many complex genetic diseases has largely motivated the development of precision medicine during the past decade. As mentioned in the paper, "a recent large-scale benchmark study, on the reliability of many common methods for observational studies in health care, found that only about 50% of the 95% intervals cover the truth" (Schuemie et al., 2020) [9]. I suspect that many of the failures might be attributed to an unrepresentative sample from a heterogeneous population.
4. There is nothing wrong with p-values, but we need to understand how to interpret them and what can go wrong. The definition of the p-value is clear: the probability of observing a test statistic as extreme as the observed one if the null hypothesis is true. However, this is not enough. There are some assumptions we must make to compute p-values. For example, there must be an underlying model, such as a linear model or normal model or binomial model. There are many other things related to that model, including randomness, correlation, no missing data, and no response bias. I think that the most critical component for statistical inference is the statistical model. Most existing inference procedures are based on a given statistical model. Where do we get the model? What happens when the model is wrong? Did we consider the model uncertainty? Nowadays, very few programs teach how to do exploratory data analysis to build a model from the observed data as John Tukey suggested (Tukey, 1977) [11]. One of the most quoted aphorisms in statistics is "All models are wrong, but

☆Main article: https://doi.org/10.51387/23-NEJSDS6.

some are useful" by George Box. As Judea Pearl commented in his blog (http://causality.cs.ucla.edu/blog/), this aphorism is painfully true but hardly useful, and it does not give us any clue as to what makes one model more useful than another. In fact, the current dilemma of statistics as a field is that most powerful statistical models for big data, such as boosting and deep learning, are invented by researchers from other fields.

5. One purpose of this paper is to be radical and encourage reactions and discussion on how we can ensure quality in statistics. This paper definitely makes a reader ponder these questions, regardless of whether they agree with the author's views.

DETAILED DESCRIPTIONS

(1) Sample Size. To further illustrate the importance of a large enough effective sample size for a heterogeneous population, let us consider a simulated example, where the underlying population is a Gaussian mixture: $0.99N(0,1) + 0.01N(5,1)$. Let μ denote the mean of the mixture distribution, and suppose that one is interested in testing the hypothesis $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ at a significance level of 0.05. Since the proportion of the second component of the mixture distribution is small, it is likely missed if the sample size n is small. When there are only samples drawn from the first component, it is known that the one-sample t-test is uniformly most powerful for the hypotheses. Figure 1(a) shows the power of the t-test for $n = 10, 20, ..., 100$, where the power at each point was estimated by the proportion of the p-values that were less than 0.05. Figure 1(a) shows a counterintuitive phenomenon: The existence of the second component of the mixture distribution tends to strengthen our confidence for accepting the null hypothesis $H_0 : \mu = 0$ when $n \leq 30$. When the sample size is small, occasional samples from the second component modify the left and right tails of the test statistic's distribution in an asymmetric way, leading to this counterintuitive phenomenon.

Figure 1(b) shows that although the t-test is sub-optimal (or theoretically wrong) for the hypothesis, it can still reach a limiting power of 1 as the sample size $n \to \infty$. However, this limit is reached at an extremely large sample size of $n \approx 10000$. It is obvious that when the sample size is reasonably large, the two components of the distribution can be well estimated using, for example, the EM algorithm (Dempster et al., 1977) [2] or the data augmentation algorithm (Tanner and Wong, 1987) [10]. This estimate can help us to understand more about the generating mechanism of the data and lead to more interpretations about the data. To conclude this discussion, I have three suggestions based on the Gaussian mixture example:

- The sample size should be reported along with the p-value. A small sample size might be able to remind people about possible unrepresentativeness of the sample for the underlying population.

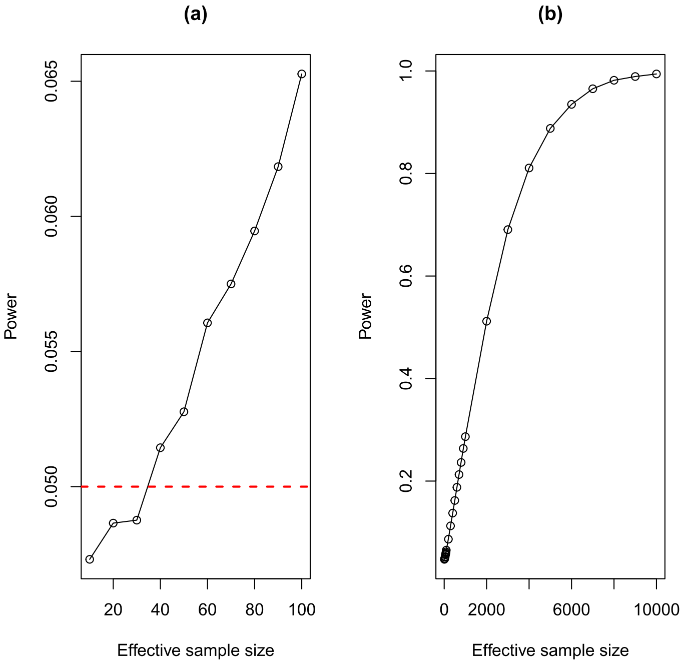

Figure 1: Power of the one sample t-test for the mean of a Gaussian mixture distribution, where the power at each point was estimated based on 105 independent tests. Figure (a) at the left is a zoomed-in version of Figure (b) at the right.

- Data integration (with an appropriate meta-analysis approach) can be an effective way to "increase" the effective sample size and thus the reliability of scientific studies.
- With big data, more statistical research can be done for interpreting the underlying population, which is closely related to popular research topics of graphical modeling and causal inference. In return, this will help researchers improve the reliability of their studies.

(2) Doubling variance. The author suggests doubling variance as a way to guard against possible local misspecification of missing data mechanisms. I have a lot of concerns about this method, which seems ad hoc and arbitrary. First, the paper does not provide any systematic guidance to when a user should use this proposal. Second, can we always ensure the desirable coverage when adopting this strategy? I would like to suggest using proven mathematical bounds to derive confidence intervals or perform hypothesis testing under some complex settings. In fact, both ideas are similar in some sense. I will use a classical statistical problem, the Behrens-Fisher Problem (Scheffé, 1970) [7], to illustrate this main idea. Suppose independent samples $X_1, \ldots, X_{n1} \sim N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{n2} \sim N(\mu_2, \sigma_2^2)$ are available. The parameter of interest is $\Delta = \mu_1 - \mu_2$. When σ_1 and σ_2 are known or proportional to each other, this is a standard problem. For the general case, there is no simple solution. Hsu (1938) [3] and Scheffé (1970) [7] provided beautiful answers for this question using probability bounds. Specifically, define $f(\sigma_1, \sigma_2) = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. We have $\frac{\bar{X} - \bar{Y} - \Delta}{f(\sigma_1, \sigma_2)} = Z_1(\xi)$, where $Z_1(\xi) = \frac{U_1}{\{\xi U_{21}^2 + (1-\xi) U_{22}^2\}^{1/2}}$. Here, $U_1 \sim N(0,1)$ and $(n_k - 1)U_{2k}^2 \sim \chi^2_{n_k - 1}$, $k = 1, 2$, are independent, and $\xi = (1 + \frac{n_1 \sigma_1^2}{n_2 \sigma_2^2})^{-1}$. Note that ξ takes values in (0, 1). Hsu (1938) [3] showed that $Z_1{*} \sim t_{\min(n_1, n_2) - 1}$ is stochastically fatter than $Z_1(\xi)$ for all ξ. Hence, a confidence interval based on $Z_1{*}$ is always a valid interval with correct coverage. This is a remarkable analysis with mathematical rigor. For more details on the analysis of the Behrens-Fisher Problem, see Martin and Liu (2015) [4, 5, 6].

SOME MORE COMMENTS

Overall. This paper covers a lot of ground, so when reading it the first time, it feels scattered. For example, when finishing Section 2, there is a jump from doubling variance in multiple imputation problems to a CarTalk radio show. This could possibly be fixed by either giving Section 3 a brief introduction, or by giving Section 2 a brief summary at the end (with a transition into Section 3). Theorem 1 is a new and interesting take on Bayes' Theorem that provides a quick way for people who are not well-versed in statistics to estimate the positive predictive value.

Section 2. The article anticipates common criticisms (especially regarding Section 2 with doubling the variance) and responds to them. In this sense, the author is aware of the reaction to some of his ideas, and he incorporates these points of view into the paper, providing a more well-rounded article. I am surprised that Section 2 (and the rest of the paper, for that matter) does not mention the value of a carefully designed experiment. For observational studies, I can see the value in doubling the variance to make the results robust to flaws in the dataset or the assumptions made in analysis. The reference Schuemie et. al (2020) [8] on Page 4 is a great example of how observational studies can produce flawed results, but Schuemie et. al. (2020) [8] appears to only consider observational studies. If an experiment is carefully designed (e.g., D-optimal, uniform, 2^{k-p} fractional factorial, . . . etc) then we should have higher quality results. Moreover, this underscores the larger problem, which is that bad data can lead to bad (or at least, easily misinterpreted) results. The author is focused on fixing the analysis after the data have already been collected by using an ad hoc procedure (doubling the variance). Another path to providing results that the public can trust is to collect data in a more intelligent way, addressing the problem before analysis even begins.

Section 3. I agree with the idea behind Section 3. The Dirtified Bayes' Theorem provides a quick shortcut that most people could use to get a close approximation to the posterior probability (i.e., the positive predictive value). It is helpful to have shortcuts, as long as we are aware of their error bounds and limitations. Perhaps PC2 (Principled Corner-Cutting) should be mentioned earlier in the section (maybe in an introduction to the section; see my earlier point), and then the article can move into CarTalk and Dirtified Bayes.

Section 4. The majority of the discussion is about an example that even the author thinks is impractical (i.e., deducting salary in academia based on the percentage of "wrong" results). This section would be better served by providing more suggestions about how we can incentivize self-quality controls in institutions, such as the discussion in paragraph 3 on page 20. I do not think that Section 4 is particularly radical. This is not a bad thing, and the author acknowledges it throughout the article. To me, it just seems like common ethical practice to self-scrutinze your own work. However, I find it odd that the value of peer review is not mentioned. In academia, most scientific journals are peer-reviewed. Is this implying that the peer review process is not enough? I understand that the author of an article is the one who is most familiar with their methodology, but surely there is some value to peer review? I think the more "radical" idea is to provide incentives for this self-scrutiny (Section 4.2). For any organization, measuring the amount of "wrong" results that a person produces would take an inordinate amount of time (or, as the author puts it, it is an NP-hard task). Moreover, it is not always a trivial task to determine if one's work is "wrong" or "disproved" in the future. If my 95% CI for μ is (3.1, 3.2) and another study reports that $\mu \approx 3.3 \pm 0.05$, am I wrong, or is the other study wrong? (Or, even worse, are we both wrong?) Did both studies use identical methodologies? I understand that the author does not want to implement such a radical idea, but he should emphasize the difficulty of going back and determining if one's work was correct.

Section 5. I agree with the ideas presented in Section 5. We should start teaching children about the concept of variation (and distributions) earlier. People should understand how to think in terms of distributions (e.g., the five-number summary at least) instead of assuming that the world is deterministic. I feel that in statistical consulting, the largest barriers occur when clients are unfamiliar with the concept of a distribution. Additionally, if more people were comfortable with the concepts of probability and variation, we would not have to be as concerned with the misinterpretation of scientific results. Of all the suggestions in the paper, I feel this is the one that should be of the highest priority.

ACKNOWLEDGEMENTS

I would like to thank my Purdue colleagues for their insightful comments on my review; specifically, Drs. Faming Liang, Chuanhai Liu, Arman Sabbaghi, Xiao Wang, Tim Keaton and Nicholas Rios.

Accepted 7 September 2022

REFERENCES

[1] Allison, K. H. and Sledge, G. W. (2014). Heterogeneity and cancer. *Oncology* **28**(9) 772–772.

[2] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1) 1–22. MR0501537

[3] Hsu, P. (1938). Contribution to the theory of" Student's" t-test as applied to the problem of two samples. *Statistical Research Memoirs*.

[4] Martin, R. and Liu, C. (2015). Conditional inferential models: combining information for prior-free probabilistic inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(1) 195–217. https://doi.org/10.1111/rssb.12070. MR3299405

[5] Martin, R. and Liu, C. (2015) *Inferential models: reasoning with uncertainty* **145**. CRC Press. MR3618727

[6] Martin, R. and Liu, C. (2015). Marginal inferential models: prior-free probabilistic inference on interest parameters. *Journal of the American Statistical Association* **110**(512) 1621–1631. https://doi.org/10.1080/01621459.2014.985827. MR3449059

[7] Scheffé, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association* **65**(332) 1501–1508. MR0273732

[8] Schuemie, M. J., Cepeda, M. S., Suchard, M. A., Yang, J., Tian, Y., Schuler, A., Ryan, P. B., Madigan, D. and Hripcsak, G. (2020). How confident are we about observational findings in healthcare: a benchmark study. *Harvard data science review* **2**(1).

[9] Schuemie, M. J., Ryan, P. B., Pratt, N., Chen, R., You, S. C., Krumholz, H. M., Madigan, D., Hripcsak, G. and Suchard, M. A. (2020). Principles of large-scale evidence generation and evaluation across a network of databases (LEGEND). *Journal of the American Medical Informatics Association* **27**(8) 1331–1337.

[10] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association* **82**(398) 528–540. MR0898357

[11] Tukey, J. W. (1977) *Exploratory data analysis* **2**. Reading, MA.

Dennis K.J. Lin. Purdue University, 150 N. University St, West Lafayette, IN 47907, USA.
E-mail address: dkjlin@purdue.edu