The New England Journal of Statistics in Data Science Volume 1, 26–28 (2023)
DOI: https://doi.org/10.51387/22-NEJSDS6B

Comment on "Double Your Variance, Dirtify Your Bayes, Devour Your Pufferfish, and Draw Your Kidstogram," by Xiao-Li Meng[☆]

Thomas R. JUNK

Abstract

This contribution is a series of comments on Prof. Xiao-Li Meng's article, "Double Your Variance, Dirtify Your Bayes, Devour Your Pufferfish, and Draw Your Kidstogram". Prof. Meng's article offers some radical proposals and not-so-radical proposals to improve the quality of statistical inference used in the sciences and also to extend distributional thinking to early education. Discussions and alternative proposals are presented.

KEYWORDS AND PHRASES: Radical, Proposal, Comment.

Professor Meng's contribution, "Double Your Variance, Dirtify Your Bayes, Devour Your Pufferfish, and Draw Your Kidstogram" provides quite a lot of provocative food for thought. Given the current concern over replicability of experimental results, it is a good idea to consider all manner of solutions moving forwards, including radical ideas. Of course, such ideas need to be scrutinized and reviewed to evaluate whether they would produce the desired solutions without incurring undesirable side effects. Many of the proposals are well thought-out and not even radical. The bulleted list of Section 4.1 is inspired and it needs to be taped to the door of every experimental researcher's office. The formulation of these questions in the past tense implies their being asked too late, however. In my own field of experimental particle physics, we ask investigators to contact statistics experts before they start their analysis work, and before they even design their experiment, if possible. That way, the experiment and the analysis can be designed so that honest, reassuring answers to these questions can be given when the results are ready. Of course, publishing this list of questions here aids in the advance preparation of investigators reading the article.

Regarding the discussion of expiration ("best by", "sell by") dates, these time intervals are largely in the hands of the food producers, and they usually have to do with food quality rather than safety. The expiration intervals are thus optimizable. Prof. Meng points out that marketing reasons encourage longer quoted freshness intervals, but that there is a "price for overdoing it", since customers who eat a stale product may think that's just how it tastes and they may therefore not purchase the product again. A short interval may cause customers to throw away food and buy more, angering some customers but increasing the revenue from others. If the shelf life is too short, supermarkets won't stock the item and customers won't buy it. Of course, the optimization goes beyond just picking an interval of time for the label – it also involves making products truly more shelf stable. These improvements are often a good thing for everyone, unless the changes to the product to lengthen its lifetime adversely impact its quality, taste, price or other factors. This discussion isn't totally beside the point of statistical inference, as improving the quality of statistical analyses is usually much preferred over simply adjusting the labeling after the fact. It is important to affix an honest label to a product, however it was produced.

Along these lines, I am concerned about the proposal to double the variance of reported results. Prof. Meng mentions that results that are used as inputs to other statistical analyses will cause biases and underestimated final uncertainties if their quoted variances do not represent the true variation, or at least the best estimate thereof. In many fields of study, it is impossible to determine in advance whether a result will be the "last" one in a chain of reasoning, or if it will be used as input to another study.

Doubling the variance can also expand the edge of a confidence interval beyond boundary that is known a priori. For example, the expected fraction of voters for a particular candidate in a national election cannot be negative, nor can it exceed 100%. What is the meaning of $10\% \pm 20\%$ in this case?

One justification for the factor of two comes from covering omitted covariance in the case of the sum of two estimates, as seen in Eq. 2.1. But if n terms are summed instead of two, the bound becomes n times the naive variance and not two. Another justification given for the factor of two is that it is the smallest among non-unity integer multipliers. Of course, there is no requirement that a conservative factor is an integer, or that it is small. It is too small in some cases and too large in others.

[☆]Main article: https://doi.org/10.51387/22-NEJSDS6.

A real concern with the state of non-replicable results, however, is the more common omission of estimations of systematic uncertainty. Covering unaccounted sources of bias by doubling the variance of the known sources is unlikely to succeed in getting the "right" answer for the variance to quote. Furthermore, some measurements may in fact not be plagued by unknown biases, and the statistical variations may be well understood. In this case, too, doubling the variance will produce misleading results. As an example, signal processing relies on an understanding of the signal-to-noise ratio (SNR). The noise in a communication channel can usually be easily measured by examining the data from it when no signal is present. Doubling the variance of the measured noise will produce a SNR that is a factor of $\sqrt{2}$ too small. This has impacts on the compressibility of data, the number of bits needed to store digital data, and thresholds needed to select signals above noisy backgrounds. This is an example of the result of one analysis feeding into another and not the end result of a long chain, though. But where one draws the line and applies a factor of two, which must be documented and undone where appropriate, is not well defined.

Some variances are well known and have real-world impacts. For example, the cone of uncertainty in hurricane center path projections is based on a large ensemble of previous hurricane paths, with coverage set at two-thirds (https://www.nhc.noaa.gov/aboutcone.shtml). Doubling the variance of hurricane forecast paths will dilute their impact, as the conventional coverage probabilities are understood as part of the planning process. P values will no longer be uniformly distributed between 0 and 1 under the null hypothesis if a doubled variance is used as part of the distribution of the test statistic to compute them.

In the Car Talk problem, Ray mis-states the answer: "So your chances of actually having it ... are one in 51" when in fact only an upper bound has been calculated. A simple reason can be explained because the prevalence p is not known; only an upper bound is given. The unknown false negative rate also does not change the upper bound. The fact that p is at most 0.1% is easy to explain, even in the short airtime allowed in Car Talk, by adding the words "less than" in the right places. If the estimate of the probability of having the disease given a positive test were used to formulate a policy or a treatment regimen, we might get the wrong answer with a definite number instead of an inequality. The difference between 0% and 2% is big enough to matter, especially if the disease is fatal. I do appreciate the detailed discussion of this matter in Prof. Meng's paper, but really, when approximations are made, one must be aware of the full calculation one is approximating, if only to determine whether the example is in the domain of applicability. It is very common for practitioners to follow an example strategy that has a corner-cutting approximation, not realizing what the full calculation ought to be, but taking the approximation as the recommended method in all similar cases. These practitioners may be confused when the approximation breaks down, causing loss of coverage in results. Knowing the better, if more cumbersome, way of arriving at an answer allows an investigator to construct a new approximation or at least consider the one being used. In the Car Talk case, identifying the missing pieces and treating the result as an inequality is rather little extra work, and is even more enlightening and instructional.

Introducing an economic incentive for investigators to produce replicable results is an interesting, curious proposal. Tying the salary reduction to α however, seems to incentivize the wrong thing. It is not only the state of being wrong that incurs a penalty under this proposal, but being wrong and being unsure of the result has a higher penalty than being wrong and being more confident. At 90% CL, an investigator ought to produce results that do not contain the true value in the confidence interval 10% of the time, resulting in a disaster for that investigator's compensation. But 90% CL results do have value, especially when labeled properly, and we do not want to discourage properly labeled results.

Prof. Meng does allude to the fact that such an "incentive system can and will induce more serious gaming behaviors or even fraudulent manipulations." It is an entertaining exercise to think of some of these behaviors and manipulations, some of which may be difficult to detect and others which may appear to those involved to be well-intentioned. It is all too easy to circumvent the financial incentive. Faculty could be supported with grant money or with private donations, and all funding sources would have to cooperate. Departments could award raises based on unrelated criteria, consciously or unconsciously filling in the penalties that some may think have been arbitrarily applied. This financial penalty also incentivizes investigators to prevent, obstruct, or simply delay replication studies, and to defend non-replicated results vigorously. Who decides when a result has been refuted? Must the original experimenter concede when another experiment comes along? Must the replication study be more definitive than the original? After all, absence of evidence of an effect is not necessarily evidence of absence, especially if the replication attempt is not as sensitive to the claimed effect as the original study. If there is tension between different experimental outcomes, how is it decided which one is right?

There is no additional incentive for investigators to attempt to replicate other studies in this proposal. If anything, reducing the pay of an investigator is a disincentive for would-be replicators, as a salary reduction creates more animosity than the refutation of a result by itself. We may end up with a cooperative solution to the Prisoner's Dilemma, in which no one is willing to attempt replications of other investigators' work, in order to avoid penalties as professional courtesy. On a smaller scale, investigators may agree not to attempt a replication of a study if a reciprocal agreement can be struck. These problems may be better addressed by incentivizing replication studies by funding them and citing them, rather than penalizing the original work.

I applaud Prof. Meng's proposals to increase statistical literacy by improving early childhood education. Distributional thinking is important and aspects of it are not even that hard to teach and grasp in elementary school. Histograms are usually introduced around the sixth grade in the United States, though students are certainly exposed to variations in measured quantities before then. The usual example I remember from my early education was distributions of student heights, expressed in various ways. The mechanics of producing such a kidstogram involve a lot of work: a student must choose a variable to measure and a sample in which to measure it. Axis ranges, bins, axis titles, colors, symbols and spacing must be chosen as well, and also whether or not to include error bars, which are frequently omitted but which are sometimes necessary to tell a complete story. These aspects are certainly fun for some (but possibly not all) students, and the experiment-design aspects, such as choosing the variable and the sample, are very instructional but can be pre-chosen for the youngest students.

Getting good at the mechanics of making a plot is only the beginning of the adventure towards expertise in using data scientifically. Some of the choices one must make when producing a graph are good to introduce to students in a concrete way, so that when they see a published graph, they know that this is not the only way to present the data and in fact the presentation may distort or hide features of the data. Histogramming is already taught in schools, but thinking critically about data quality and presentation is usually not given enough emphasis, or at least it was not when I was in school. Simply choosing a different binning for a histogram can alter the story it tells. It might be good to expose students to data handling, preparation, and presentation tools so that they can see how these steps can be used to illuminate or mislead. Software can ease the chore of making artful hand-drawn histograms when the point may be to explore a space of different options for data selection and display. The biggest risk is that young students may stop paying attention if an exercise lasts too long and the rewards are spaced too far apart. Art-project style exercises are good for kids to stay on task because they can see their creations taking shape before their eyes. But to drive home more nuanced critical-thinking lessons, a computer exercise may be more engaging. These may be more suited for older students.

I would value teaching strategies that encourage improving data quality, and also techniques for evaluating the quality of data, over data presentation methods, even if they must be packaged in a form that appeals to youngsters. Even the youngest students can appreciate the distortions introduced in the "telephone game", which points out how easy it is for uncontrolled data transmission and storytelling to affect the integrity of the data-collection process. Of course these goals are not exclusive – drawing a good histogram is the end product of a much more involved exercise.

One precursor to troubles in data analyses is the time it takes to go from a question to an answer. Sometimes things are forgotten along the way, causing errors, or shortcuts are taken and later forgotten. Keeping track of all the necessary details, and realizing that sometimes a small detail may have a big effect on the result, is an important lesson.

The idea of sampling to a foregone conclusion can also be easily conveyed in school. In fact, nearly all children have used the strategy of continuing a game or contest after losing a single round with a challenge of "best two out of three". One can often keep playing until one gets the desired outcome, at least if one player gets to decide when to stop the process at the expense of the other. This is an idea that can be introduced early and which provides contact with children's everyday lives.

One word we use very commonly in data selection is "sculpting" of distributions. Kids may find this analogy easy to grasp, though it must be introduced in a way that holds their attention. When a piece of clay is sculpted, the result of each cut is immediately visible. One can incorporate this in the kidstograms fairly easily, by dividing a sample into two subsamples, and coloring the histogram as a stacked mixture. "How many times a day do you sharpen a pencil?" can have two components – days with tests, and days without tests, for example. This tells a story that goes beyond mere observation, and starts touching on the mechanics of performing controlled experiments. Gently introducing the basics of the scientific method is important, and showing students how to present data is but one piece of such an introduction.

ACKNOWLEDGEMENTS

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

Accepted 7 September 2022

Thomas R. Junk. Fermi National Accelerator Laboratory, MS 220, P. O. Box 500, Batavia, IL 60510 USA.
E-mail address: trj@fnal.gov