# Double Your Variance, Dirtify Your Bayes, Devour Your Pufferfish, and Draw Your Kidstogram*

Xiao-Li Meng
Department of Statistics, Harvard University

August 1, 2021

**Abstract**

This article expands upon my presentation to the panel on "The Radical Prescription for Change" at the 2017 ASA (American Statistical Association) symposium on *A World Beyond* $p < 0.05$. It emphasizes that, to greatly enhance the reliability of—and hence public trust in—statistical and data scientific findings, we need to take a holistic approach. We need to lead by example, incentivize study quality, and inoculate future generations with profound appreciations for the world of uncertainty and the uncertainty world. The four "radical" proposals in the title—with all their inherent defects and trade-offs—are designed to provoke reactions and actions. First, research methodologies are trustworthy only if they deliver what they promise, even if this means that they have to be overly protective, a necessary trade-off for practicing quality-guaranteed statistics. This guiding principle may compel us to doubling variance in some situations, a strategy that also coincides with the call to raise the bar from $p < 0.05$ to $p < 0.005$ (Benjamin et al., 2018). Second, teaching principled practicality or corner-cutting is a promising strategy to enhance the scientific community's as well as the general public's ability to spot—and hence to deter—flawed arguments or findings. A remarkable quick-and-dirty Bayes formula for rare events, which simply divides the prevalence by the sum of the prevalence and the false positive rate (or the total error rate), as featured by the popular radio show Car Talk, illustrates the effectiveness of this strategy. Third, it should be a routine mental exercise to put ourselves in the shoes of those who would be affected by our research finding, in order to combat the tendency of rushing to conclusions or overstating confidence in our findings. A pufferfish/selfish test can serve as an effective reminder, and can help to institute the mantra "Thou shalt not sell what thou refuseth to buy" as the most basic professional decency. Considering personal stakes in our statistical endeavors also points to the concept of *behavioral statistics*, in the spirit of behavioral economics. Fourth, the current mathematical education paradigm that puts "deterministic first, stochastic second" is likely responsible for the general difficulties with reasoning under uncertainty, a situation that can be improved by introducing the concept of histogram, or rather *kidstogram*, as early as the concept of counting.

*Keywords*: Behavioral Statistics, K-12 Mathematical Education; Outerval; $p$-value; Quick-and-dirty Bayes Theorem; Research Replicability and Reliability; Principled Corner Cutting (PC2); Quality-guaranteed Statistics; Selfish Test; Soft Elimination.

*I thank Ron Wasserstein for inviting me to be radical back in 2017, and *New England Journal of Statistics in Data Science* for providing me with this new forum to expand my thoughts on paper, especially after three years of adventure in the world of data science as the founding Editor-in-Chief of *Harvard Data Science Review*. I also thank the audiences of an FDA (Food and Drug Administration in the US) seminar and of WHOS-PSI III for their encouragement and feedback on a presentation based on this article, and Kai Zhang for his great help in making Figures 1 and 2, as well as for double checking my algebra. I am deeply grateful to a dozen colleagues and friends who have helped to rationalize or radicalize my proposals to ensure a healthy balance: Joe Blitzstein, Radu Craiu, Christine Franklin, Robert Gibbons, Pierre Jacob, Thomas Junk, Eric Kolaczyk, Nicole Lazar, Louis Lyons, Natesh Pillai, Nathan Sanders, and Lance Waller. Of course, readers should blame any irritation on my inability in practicing what I preach (e.g., communicating effectively) and in optimizing under constraints (e.g., optimally weighting and contrasting competing perspectives).

# 1 Be As Radical As You Wish ...

> "*Unlike the ASA Statement, which attempted to be a consensus document, we are not trying to build consensus at this symposium. We are trying to effect change, even though we might not all agree on what change. So be as radical as you wish.*"
>
> Ronald Wasserstein, Executive Director, American Statistical Association (ASA)

The ASA Statement referred to in Ron's invitation is perhaps the most widely debated—or at least discussed—statement ASA ever issued: The ASA Statement on $p$-values (see Wasserstein and Lazar, 2016). I felt lucky that I was invited to be radical rather than consensual, for I did not envy at all the task faced by the authors of the $p$-value statement. Ron, however, did include some fine print with his encouragement: "If your radical prescription involves closing down the ASA or firing its executive director, I would appreciate two weeks notice."

Whereas the thought of closing down or firing was not on my mind, *elimination* was, when I started to contemplate which radical track I should follow. A symposium on $p$-value obviously would have many talks on $p$-value, so I could at least avoid talking about *that*. After all, there are a good number of statisticians who would avoid the use of $p$-value, or even hypothesis testing, entirely. They would prefer estimation, whether point estimators, interval estimators, or distributional estimators (e.g., Xie and Singh, 2013).

However, there is a fundamental premise underlying hypothesis testing that no statistical inference or prediction procedure can avoid. We repeatedly emphasize to our students that, when a null hypothesis survives a statistical test, they should never declare the acceptance of the hypothesis but only that the test fails to reject it. Cynical minds may consider this careful wording is statisticians' way to cover their assets. Anyone who has a good understanding of statistical tests, however, would have to agree that this is the only logical conclusion one can reach from this test result alone. Statistical testing is about determining if there are sufficiently large discrepancies between a (null) hypothesis and the observed data, according to a pre-specified probabilistic criterion. Because no test is almighty, when a test fails to find the discrepancies in selected aspects, it says little about discrepancies in other aspects. More importantly, for any data set, we can find uncountably many models (and hypotheses) that fit the data perfectly: any zigzagging curves or surfaces that connect every observed data point cannot be rejected by any pure "goodness-of-fit" test by definition. Yet almost all such "models" would look and sound ridiculous for any purposes other than illustrating how ridiculous they are. Therefore, we can *eliminate* a hypothesis when the test finds sufficient discrepancies with the data, but we can say little about its validity otherwise. Readers who enjoy reading (and thinking) about philosophy of science may wish to look into the extensive writings on this line of thinking by Karl Propper, one of the greatest philosophers of science of the twentieth century[1].

More precisely and generally, statistical inference and prediction is a soft-elimination game, by declaring certain *a priori* permissible values of our target are no longer plausible according to a pre-specified criterion evaluated on the actual data. It is a *soft elimination* because the implausible values are not mathematically impossible, and indeed we may bring some of them back in light of new data or understanding. It is a *game* because it has—or should have—clearly stated rules, and our opponent is nature (or God or devil or whomever/whatever we should thank for challenging our intelligence and for making our profession vital).

A confidence—or posterior/fiducial—interval therefore is more about declaring that any value outside of it can be eliminated from further consideration, as a way to sharpen our inference driven by our insatiable

[1] See `https://plato.stanford.edu/entries/popper`

desire for certainty, than saying anything about the truthfulness of the values inside of it. Or, using a term attributed to John Tukey (see O'Rourke's comments on Gelman's blog about "uncertainty interval"[2]), our aim is to seek an *outerval* to eliminate implausible values as declared by our chosen criterion. We therefore might modify Gelman's proposal (in the same blog) of replacing "confidence interval" with "uncertainty interval" by adopting the less inflammatory term "plausible" for "confidence."

From this perspective, the general preference to err on over-coverage than under-coverage is not much about being conservative or liberal. Indeed, equating over-coverage with being conservative can seriously mislead ourselves. For example, Junk and Lyons (2020) pointed out that over-estimating the sensitivity of a nuisance parameter can lead to under-estimating uncertainties for parameters of interest. Rather, it is a practical way to ensure delivering on promise, a critical step for gaining general trust in statistical methodologies and for improving scientific replicability and reliability. Section 2 illustrates this point in the context of doubling variance as a small premium for insuring against hidden complications (e.g., inestimable dependence) that may render our promise misleading. Section 3 explores further the usefulness of moving away from conventional methods for the purpose of gaining practical insights that can increase the chance of detecting and preventing flawed reasoning, as illustrated by a "quick-and-dirty" Bayes formula for rare events. Section 4 and Section 5 move from methods and theorems into ethics and education, proposing respectively a not-so-radical ethical test for research confidence and a more radical pedagogical paradigm for early childhood education to teach stochastic thinking as early as the deterministic manipulations. Section 6 concludes with a call for action to ensure the vitality of statistics in the data science ecosystem.

An acknowledgement and warning before proceeding. An early draft of this article was sent to about a dozen of researchers and educators, and their reactions made it clear that my attempt at being radical is partially successful. The most diverse and strong reactions are to the suggestion on doubling variance, varying from "really nice" to "a dangerous idea". I am deeply grateful to all the previewers for their candid criticisms, inspiring insights, and reasoned reflections, which have given me much food for thought. Some of the concerns I clearly had overlooked (e.g., competitive advantages of researchers who adopt different criteria). Others reminded me once more of the importance of effective communication, especially when presenting methods with general appeals (e.g., easy to implement) but come with a long list of caveats. For example, we clearly should not double variance unnecessarily; we obviously need to worry about the negative consequences of over-assessing uncertainties; and we ought to seek methods that can do better without incurring undue cost. The need for listing such caveats tends to be more evident in their general forms and in abstraction than in specific studies, where our ability to conduct disinterested and critical introspection tends to be reduced by our goal-oriented passion and investment in the studies. I therefore invite readers to join me in a somewhat demanding journey as we explore together the "radical" proposals in this article: navigating between generality and particularity to inform and form a collective strategy for communicating and realizing the benefits of each proposal while containing and reducing its negative impact.

# 2 Deliver on Our Promises: Double the Variance (and Our Effort)

## 2.1 Your method is expired ...

Many products, especially for human consumption, come with an expiration date: milk, juice, canned food, medications, etc. The date is a guarantee of freshness or efficacy when the product is consumed prior to

[2] https://andrewgelman.com/2010/12/21/lets_say_uncert/

it. To provide such guarantees for (nearly) all individual products, the declared expiration date cannot be some average of individual expiration dates, but rather a (statistically) safe lower bound. We therefore may still consume a product after its expiration date but at our own risk. Even if some producers may desire to declare longer shelf-life for marketing reasons, they (should) understand well the price of overdoing it. A few bad cases may ruin the public's trust in a producer, and hence it is in the latter's best interest to be better safe than sorry.

The same principle should apply to research methodologies and receipts, at least to those that are for general consumption. When we invoke an interval procedure with a declared 95% confidence or posterior probability or whatever term we use, we are telling ourselves, and everyone else, that the values excluded by our procedure account for no more than 5% of the term we adopt. It is therefore safe to remove those values outside of the interval from further considerations, when we decide and declare that 5% is our threshold. That is, our trust in a procedure lies in its delivery of its promises. If we are comfortable about eliminating a state from further consideration at the 5% level, then we can accept eliminating it at any level below 5%, just as a consumer we do not mind—indeed we hope—that the actual expiration date exceeds the declared one. But we cannot, and should not, accept an error above 5%, just as we would not be happy if a glass of milk tastes sour prior to the stated expiration date (and we have stored the milk properly as instructed).

Unfortunately, much of our current practice does not come with such a seemingly minimal quality guarantee. The lack of quality by no means occurs only to estimation or hypothesis testing procedures, but their defects are most visible because many of our confidence procedures are verifiable (e.g., via simulation) to have significantly less coverage than the promised one. 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). This should be a very scary finding for anyone who cares about reliability of statistical methods—few business entities can do business as usual, if half of their products are found to be defective. This is sadly not the first time—or the last time—that such staggering quality disasters are revealed (e.g., Ioannidis, 2005; Simmons et al., 2011). Indeed, it is these types of findings that have led to general concerns of the so-called "replicability crisis", which has generated many discussions, debates, and decisions. ASA's statement on $p$-values and this subsequent symposium is one of them, so was the 2019 report on "Reproducibility and Replicability in Science" issued by the US's National Academies of Sciences, Engineering and Medicine; see the special theme collection with the same title in *Harvard Data Science Review* (*HDSR*)[3], especially the interview (Fineberg et al., 2020) and the introduction (Stodden, 2020).

As a part of our effort to combat such persistent problems, we need to stress constantly the need for quality control especially in situations where we tend to slip. For example, the shortage of (confidence) coverage could be due to inaccuracies in the mathematical or numerical approximations we adopt, or because of flawed applications, such as applying a procedure built for independent observations to cases where the dependence is not negligible. In practice, using approximations in methods or modeling is the rule rather than the exception. But this fact does not justify ignoring quality assurance. A pharmaceutical company cannot excuse itself for providing no expiration date or giving a misleading one simply because it cannot determine accurately the date of reduced efficacy. Rather, because of the approximate nature of inference or prediction, quality assurance at every step is critical for ensuring the overall reliability of our findings, especially in view

[3]See `https://hdsr.mitpress.mit.edu/reproducibilityandreplicability`. For full disclosure, I have served as Founding Editor-in-Chief of *HDSR* since July 2018

of the general tendency of rushing to conclusions induced by our current systems of incentives (see Section 4).

A reader may question that, given all the approximations we make and all the uncertainties we face, does it really make sense to worry about, say, a nominal 95% confidence interval procedure having actual coverage of 92%? After all, the whole concept of coverage is a thought experiment over some idealized set of hypothetical replications we conceive to be relevant. For any particular application, if we can be sure that this under-coverage is the only leeway we allow ourselves, then it may well be counterproductive to worry about a small deterioration of quality. The trouble is that, exactly because we necessarily make all kinds of approximations in an inference or prediction process, errors can accumulate in ways far more damaging than we expect or even understand, leaving us in a vulnerable position to say the least. (The phenomenon that a seemingly tiny data defect correlation (e.g., 0.005) due to selection bias can cause over 95% loss of effective sample size is such a vivid example (Meng, 2018).) Establishing a protocol and habit of ensuring quality at every step goes a long way in reducing this vulnerability. This practice is not merely to enhance our professional ethical code. It also reveals methods that otherwise would be deemed inferior.

## 2.2 Double the variance or adding up the standard errors?

As an example, Copas and Eguchi (2005) proposed to double the variance as a way to guard against possible local misspecifications of missing data mechanisms. Under the conventional mindset of "getting it right on average at least approximately", this suggestion may sound very conservative, and indeed it can be. However, when our priority is to ensure our procedures deliver what they promise, we may be much more willing to pay a premium to insure against disastrous violation. A recent study of multiple imputation (MI) inference under uncongeniality demonstrates this preference well (Xie and Meng, 2017), as summarized in Section 2.6. It shows how the strategy of doubling variance provides an extremely simply practical solution to a long standing challenge of *uncongeniality* due to the incompatibility between the imputer's (Bayes) model and user's (frequentist) procedure (Meng, 1994a). Those readers who do not need to be convinced by a technical illustration can skip that section. Nevertheless, the specific argument that led to the "doubling variance" proposal there is worthy of highlighting because it comes in rather handy in some situations, such as when we have (reliable) estimates of variances of individual components/estimates, but not of their covariances. Such kind of problems with missing correlation information have arisen in many areas and hence are continuously being researched in multiple disciplines; see Koch (2021) for a most recent work in physics.

As a specific case, suppose our confidence procedure (or hypothesis testing) calls for the evaluation or estimation of $\mathrm{V}(\hat{\theta}_1 + \hat{\theta}_2)$, but we have information only to estimate the individual $\mathrm{V}(\hat{\theta}_1)$ and $\mathrm{V}(\hat{\theta}_2)$. Can we still guarantee to deliver at least the declared coverage probability or not to exceed the stated Type-I error in a meaningful way (that is, not to use trivial procedures such as employing the entire line as our confidence interval)? The not uncommon practice of pretending zero correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$—and hence (erroneously) using $T = \mathrm{V}(\hat{\theta}_1) + \mathrm{V}(\hat{\theta}_2)$ for $\mathrm{V}(\hat{\theta}_1 + \hat{\theta}_2)$, clearly does not do the job. But doubling $T$ will do because

$$\mathrm{V}(\hat{\theta}_1 + \hat{\theta}_2) = \mathrm{V}(\hat{\theta}_1) + \mathrm{V}(\hat{\theta}_2) + 2\mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_2) \le 2T \tag{2.1}$$

regardless of the relationship between $\hat{\theta}_1$ and $\hat{\theta}_2$, a simple consequence of $\mathrm{V}(\hat{\theta}_1 - \hat{\theta}_2) \ge 0$. This inequality holds more generally for multivariate $\theta$, in terms of semi-positive definitiveness.

Interestingly, for a univariate $\theta$, a sharper bound is obtained by adding standard errors (SE) instead of

variances. That is, regardless of the relationship between $\hat{\theta}_1$ and $\hat{\theta}_2$,

$$\mathrm{SE}(\hat{\theta}_1 + \hat{\theta}_2) \leq \mathrm{SE}(\hat{\theta}_1) + \mathrm{SE}(\hat{\theta}_2), \tag{2.2}$$

which is a consequence of $|\mathrm{Corr}(\hat{\theta}_1, \hat{\theta}_2)| \leq 1$. When $\mathrm{V}(\hat{\theta}_i)(i = 1, 2)$ can be regarded as known (e.g., either can be calculated or approximated to desired accuracy), we clearly should use (2.2) instead of (2.1) because there is no reason to pay extra premium when we do not have to. However, when $\mathrm{V}(\hat{\theta}_i)(i = 1, 2)$ themselves need to be estimated or approximated, as typical in practice, if our goal is to reduce false quality assurance, then it is justifiable to seek additional protection by using a mathematically more generous bound. Hence we may prefer (2.1) over (2.2), when constructing a confidence interval (to ensure the stated coverage is no lower than declared) or hypothesis testing (to ensure the stated Type-I error rate is not exceeded).

## 2.3 Let's practice quality-guaranteed statistics

The emphasis on extra protection against exceeding stated Type-I errors naturally would lead to the question "But what about Type-II errors?" This question must be on many readers' minds. Indeed it has been on mine too since I embarked on this quest to "double the variance, double the fun(d)"[4]. Unlike the debate on $p$-values, there is a crisp answer—or rather question—here: *What is the stated criterion?* In the context of hypothesis testing, the stated criterion (in the Neyman-Person setup) has almost always been to control Type-I error first, and then to seek the most powerful test, that is, to minimize Type-II error. Controlling Type-I error, i.e., false positive rate, obviously cannot be the only criterion, but it is also well-understood that there is no free lunch—it is mathematically impossible to simultaneously minimize both false positive and false negative errors in general. We therefore must choose or compromise (e.g., using the *total error*, as in Section 3). When our stated and advertised criterion is that "statistically significant at the level 0.05", we must first deliver that promise. Indeed, we all have been taught that it is not meaningful to compare the powers of two tests without first equating/controlling their Type-I errors, for the obvious reasons.

If our stated criterion is to control Type-II error first, e.g., the test must be at least 80% powerful under a specified alternative hypothesis, then indeed we may have to find a (non-trivial) lower bound of the variance or standard error (such as $\mathrm{SE}(\hat{\theta}_1 + \hat{\theta}_2) \geq |\mathrm{SE}(\hat{\theta}_1) - \mathrm{SE}(\hat{\theta}_2)|$ when $\mathrm{SE}(\hat{\theta}_1) \neq \mathrm{SE}(\hat{\theta}_2)$). Obviously, such a procedure may have a (much) larger Type-I error than when we actually know the standard error, say, the value of $\mathrm{SE}(\hat{\theta}_1 + \hat{\theta}_2)$. If we want to ensure that our desire for 80% power will not cause unacceptably large Type-I error, then we need to specify our tolerance level for it. In general, we need to be clear about any criterion that will become a part of our guarantee. For example, we may want our procedure for the null hypothesis, $\theta = 0$, to have (I) at last 70% power when $\theta > 1$ and (II) but no more than 15% Type-I error rate. To practice quality-guaranteed statistics means that we must either deliver a procedure with these two properties guaranteed or state the properties the procedure actually delivers.

For instance, it is entirely possible that we are unable to find any procedure that we can prove to possess both (I) and (II), but we can guarantee (I) if we relax (II) to "no more than 30% Type-I error". Whereas this new threshold doubles the tolerable Type-I error rate as we stated, we have no choice but to honestly report this higher false positive rate, if we insist on guaranteeing 80% power. Or we may choose to relax (I) from 80% to 70% in order to protect (II). If we find neither is acceptable, we also have the option, at least in principle, to work harder to collect more data and information to achieve our original goal. This is the case, for example, with study size determination based on power considerations (Roy et al., 2007;

[4] I, however, will not ruin any interested reader's double fun to explicate the double puns here!

Bhaumik et al., 2008, 2009; Amatya et al., 2013), where we can also derive statistically safe lower bounds by using low confidence interval end points for effective sample size that take into account considerations such as non-response rates and data defect correlations (Meng, 2018; Isakov and Kuriwaki, 2020; Bradley et al., 2021).

All these options are within the domain of practicing quality-guaranteed statistics and science, and hence each of them can help to enhance replicability of statistical and scientific studies. What is unhelpful, and typically harmful, is to deliver procedures without the stated quality guaranteed, or with wishful claims but without disclaimers. A consumer may choose to take a dietary supplement despite its disclaimer "The stated benefits have not been evaluated by FDA" on its label, but that would be the consumer's informed—though not necessarily wise—decision. In a similar vein, if we must deliver a procedure without quality guarantee, then minimally we should remind practitioners of the potential of overconfidence by providing warnings such as "The actual performance of this nominal 95% confidence interval procedure has not been established."

## 2.4 Extra protections: Chebyshev confidence and $p < 0.005$

Doubling variance, like any (practical) statistical procedure or rule of thumb, obviously is not a universal recipe. Clearly it does not necessarily deliver desired quality when the variance itself is of inferential interest, such as for assessing volatility of financial instruments. We also should not double a variance estimate when we already know it is an overestimation, such as when the imputer's model is nested within the user's model in the context of MI inference and both models are valid (Xie and Meng, 2017). The suggested doubling strategy is for deriving our *final* confidence interval or $p$-value, yet we have good reasons to believe that our variance term has not captured some major sources of uncertainty, such as from model mis-specification (e.g., very questionable assumption of independence) or from model over-fitting (e.g., due to adaptive model selection). However, as I was reminded by several of the previewers, one study's final confidence internal can be another study's input, and hence it is important to report explicitly the doubling variance strategy when adopting it. This is not an extra burden, but rather a reminder of always being transparent about the data and process that lead to our findings, unless there are legitimate proprietorial or privacy constraints

Doubling variance is an embarrassingly practical procedure to provide some extra protections in cases where it might not seem necessary initially, or where researchers have worked hard to obtain more sophisticated and hence costly remedies. It is an approximation in nature, but as John Tukey emphasized 60 years ago (Tukey, 1962), "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." Considering the case where we have good confidence in the normal approximation to the distribution of our estimator $\hat{\theta}$ as well as our variance assessment $\sigma^2 = \sigma^2(\hat{\theta})$, and hence the usual confidence interval (say) $(\hat{\theta} - 2\sigma, \hat{\theta} + 2\sigma)$ should have nearly 95% coverage, as claimed. Doubling the variance therefore seems unnecessary. However, if we double the variance, namely, we replace $\sigma$ by $\sqrt{2}\sigma$, the resulting interval $(\hat{\theta} - 2\sqrt{2}\sigma, \hat{\theta} + 2\sqrt{2}\sigma)$ would have 99.5% coverage. From an inferential point perspective, this increased confidence should be welcomed not only because it helps to guard against various approximation errors (e.g., when we replace $\sigma$ by an estimator $\hat{\sigma}$) but also because the extra protection comes with a reasonable cost, adding about 40% width. Coincidentally, for hypothesis testing, this result turns out to be practically the same as raising the bar from $p < 0.05$ to $p < 0.005$ under a normal approximation because $\Pr(|Z| \geq 2\sqrt{2}) = 0.0047$), which can be argued from several perspectives (see Benjamin et al., 2018).

More generally, by the Chebyshev inequality, if $\hat{\theta}$ is an unbiased estimator of $\theta$, then

$$\Pr\left(|\hat{\theta}-\theta| \le c\sigma\right) \ge 1-\frac{1}{c^2}. \tag{2.3}$$

Therefore, if we take $c=2$, then the usual $(\hat{\theta}-2\sigma, \hat{\theta}+2\sigma)$ would have at least 75% coverage regardless of the distribution of $\hat{\theta}$. If we double the variance, which is the same as letting $c=2\sqrt{2}$, then the guaranteed coverage, again regardless of the distribution of $\hat{\theta}$, would be $7/8=87.5\%$. Therefore, doubling the variance also provides some reasonable insurance against the assumption of normal approximation. We of course can guarantee any level $1-\alpha$ by letting $c=1/\sqrt{\alpha}$. For example, the interval $(\hat{\theta}-4.5\sigma, \hat{\theta}+4.5\sigma)$ will guarantee 95% coverage regardless of the distribution of $\hat{\theta}$ as long as it is unbiased for $\theta$, because $1/\sqrt{0.05}=4.47<4.5$.

Finally, doubling the variance also provides reasonable insurance against the bias in our estimate $\hat{\theta}$. As an illustration, suppose $\mathrm{E}[\hat{\theta}|\theta]=\theta+\beta$, where $\beta$ represents the bias. Consequently, the mean-squared error $\mathrm{MSE}(\hat{\theta})=\beta^2+\sigma^2$. Then the same Chebyshev inequality (argument) tells us that

$$\Pr\left(|\hat{\theta}-\theta| \le c\sigma\right) \ge 1-\frac{\mathrm{E}[\hat{\theta}-\theta]^2}{c^2\sigma^2}=1-\frac{1+b^2}{c^2}, \tag{2.4}$$

where $b=|\beta|/\sigma$ measures the magnitude of the bias relative to the standard deviation. By comparing (2.4) with (2.3), we see clearly how the reduction of guaranteed coverage is determined by $b$. For example, if $b=1$, that is, if 50% of our MSE is due to the bias $\beta$, then $(\hat{\theta}-2\sigma, \hat{\theta}+2\sigma)$ no longer has the guaranteed 75% coverage, but only $50\%(=1-2/2^2)$ coverage. However, if we double our variance, which means we use $c=2\sqrt{2}$, then the guaranteed coverage when $b=1$ returns to $1-(2/8)=75\%$. In general, it is easy to see that doubling variance, that is, using $c=2/\sqrt{\alpha}$, protects the claimed coverage $1-\alpha$ from having the bias as large as standard error in magnitude, that is, permitting $b$ to be as large as 1.

Interestingly, this protected range $b \le 1$ can be extended by about 20% (for $\alpha-0.05$), if we permit ourselves to adopt the normal assumption. That is, if $\hat{\theta} \sim N(\theta+\beta, \sigma^2)$, then

$$\Pr\left(|\hat{\theta}-\theta| \le c\sigma\right)=\Phi\left(c-b\right)-\Phi\left(-c-b\right).$$

Therefore, when we double the variance, the coverage becomes $\gamma=\Phi\left(\sqrt{2}c-b\right)-\Phi\left(-\sqrt{2}c-b\right)$. Hence when $c=2$, we can have $b$ up to 1.18 and still guarantee $\gamma \ge 0.95$.

## 2.5 Why doubling but not tripling or using some other multipliers?

As I reported earlier, reactions to the idea of doubling variance have been rather mixed, with several questioning the use of the factor 2, which seems ad hoc and arbitrary. In a specific context, such as given in previous sub-sections and in the next one, the theoretical justifications of doubling are as rigorous as or even more so than, say, adopting large-sample approximations, which are in routine use but without routine check on their applicability or accuracy. My colleague Joe Blitzstein also provided a compelling reason (attributed to Joe Gastwirth) for using mathematically proven bounds rather than approximations in presenting evidence: "If you are testifying as an expert witness and you've proven an upper bound on a probability $p$, you can say very confidently (no pun intended) that $p$ is less than the upper bound. It may look much weaker to the jury if you only have an approximation for $p$, when the opposing lawyer can easily question whether it is possible that the approximation is far off from the truth."

Indeed, being able to easily and persuasively communicate statistical methods to general users or findings to general audiences is a critical motivation for advocating methods such as doubling the variance. Or, as

Aaditya Ramdas commented succinctly (*personal communication*), "I like this price of 2 idea. It's not much, easy to state, and buys some nice robustness..." It is a small—indeed the smallest among integer multipliers—*price* or *premium* for insuring against defects of a procedure, and it is the easiest idea of all to convey. Ramdas also reported several cases where one achieves a safer bound by doubling the original answer (e.g., Katsevich and Ramdas, 2018). An earlier example that I encountered is in calibrating posterior predictive p-values (ppp). It is well-known that, for a frequentest p-value $p$ for testing a (null) model $M$, we have $\Pr(p \le \alpha) \le \alpha$ for any $\alpha \in [0,1]$, under $M$, assuming the test is exact. This result does not carry over to a ppp because the dependence of the posterior on data. However, one can show that $\Pr(ppp \le \alpha) \le 2\alpha$ for any $\alpha (< 0.5)$ under the Bayesian null, i.e., the prior predictive distribution under $M$ (Meng, 1994b).

In specific problems or domains, we can and should develop more sophisticated adjustment factors for procedural improvements and for validity protections. For example, Particle Data Group (PDG), an international collaborative team that publishes the *Review of Particle Physics*, advises a data-dependent inflection factor (see Particle Data Group, 2020, p. 16) for assessing measurement errors when they are suspected to be large. However, when we compare methods for reducing irreplicable and unreliable studies (or other similar general problems) at a general level, we need to consider their respective success rates in actual practice. The aforementioned "replication crisis" is far more about false positive results than false negative results (Meng, 2009b). Being overly confident perhaps is the second most likely statistical culprit for generating too many false positive results (the first being cherry picking). Statistically speaking then, doubling the variance, as a general strategy for ensuring replicability, would do more help than harm in our effort to improve replicability compared to not using it. At the same time, because it is the smallest integer as an inflation factor, it also minimizes the potential negative consequences of being too overly protective.

## 2.6 Doubling variance for combating uncongeniality in multiple imputation

MI (Rubin, 1987) was originally motivated by the fact that typical data collectors, such as the US Census Bureau, have much more information and resources to handle the missing data (e.g., due to non-responses), $Y_{\rm mis}$, than individual users of the data. The MI approach asks the data collector to impute each missing value $m$ times (e.g., $m = 10$) from an appropriately constructed imputation model. These $m$ sets of imputation are combined with the observed data $Y_{\rm obs}$ to form $m$ *completed data sets*, $D_1, \ldots, D_m$, where $D_\ell = \{Y_{\rm obs}, Y_{\rm mis}^{(\ell)}\}$, with $Y_{\rm mis}^{(\ell)}$ being the $\ell$th imputation for the missing $Y_{\rm mis}$. The users can then apply their chosen complete-data procedures to each of these $m$ data sets to obtain $m$ sets of complete-data results, say $\{\hat\theta(D_\ell), U(D_\ell), \ell = 1, \ldots, m\}$, where $\hat\theta(D)$ is users' complete-data estimator for the parameter $\theta$ and $U(D)$ is an estimator of the variance of $\hat\theta(D)$ when the users have access to the complete data $D$. These results are combined according to Rubin's combining rules (Rubin, 1987) to form the so-called MI inference. In particularly, the MI estimator is simply the average of $\hat\theta(D_\ell), \ell = 1, \ldots, m$, denoted by $\bar\theta_m$. The variance of $\bar\theta_m$ is estimated via Rubin's variance combining rule that decomposes the total variance estimate $T_m$ into the within-imputation variance $\bar U_m$ and the between-imputation variance $B_m$, which are respectively the sample average of $U(D_\ell)$ and sample variance of $\hat\theta(D_\ell)$. More precisely,

$$T_m = \bar U_m + \left(1 + \frac{1}{m}\right) B_m, \tag{2.5}$$

where the extra inflection factor $(1 + m^{-1})$ is due to the use of a finite $m$.

From early on, a key controversy about MI is the issue of *uncongeniality*, that is, the imputation model may not be compatible with the users' procedures (Meng, 1994a). A serious consequence of this unconge-

niality is that it can lead to biases in Rubin's variance combining rule (Fay, 1992; Kott, 1995), even if we assume both the imputer's model and the analyst's procedure are valid (if either of them is invalid, then it is not surprising something can go wrong). This uncongeniality is inevitable because it is not possible for the imputers to anticipate all the analyses that would be performed on the imputed data sets. Even if they could, they would not be able to construct one coherent imputation model that would be compatible with all these analysis procedures, because they may not be compatible with each other. Furthermore, the users are not given all the information about the imputation models, and indeed some of the information may be protected by confidentiality constraints, which was a part of the reason that we want the data collectors to perform the imputation in the first place. Hence the problem is unsolvable if we insist on a consistent variance estimator or exact coverage.

However, there is an extremely simple solution when we are guided by the goal of delivering quality for our confidence procedure, as discovered by Xie and Meng (2017). To see the essence of this approach most clearly, let us assume $m = \infty$ to avoid the distraction of Monte Carlo error due to finite number of imputations $m$. By the trivial decomposition $\bar{\theta}_\infty = \hat{\theta}(D) + \left[\bar{\theta}_\infty - \hat{\theta}(D)\right]$, we have

$$\mathrm{V}(\bar{\theta}_\infty) = \mathrm{V}\left[\hat{\theta}(D)\right] + \mathrm{V}\left[\bar{\theta}_\infty - \hat{\theta}(D)\right] + 2\mathrm{Cov}\left[\hat{\theta}(D), \bar{\theta}_\infty - \hat{\theta}(D)\right].$$

As shown in Xie and Meng (2017), regardless of uncongeniality, the first two terms on the right-hand side are consistently estimated by $\bar{U}_\infty$ and $B_\infty$ respectively. Under congeniality, the third cross term is zero, and hence $T_\infty$ of (2.5) (with $m = \infty$) is a consistent estimator of $\mathrm{V}(\bar{\theta}_\infty)$. Otherwise this third term can be negative or positive, and it is not possible to estimate it consistently based on $\{D_1, \ldots, D_m\}$ alone. However, because

$$2\mathrm{Cov}\left[\hat{\theta}(D), \bar{\theta}_\infty - \hat{\theta}(D)\right] \le \mathrm{V}\left[\hat{\theta}(D)\right] + \mathrm{V}\left[\bar{\theta}_\infty - \hat{\theta}(D)\right],$$

we see that if we use $2T_\infty$ as an estimator for the absolute upper bound of $\mathrm{V}[\bar{\theta}_\infty]$, we can guarantee the resulting confidence interval to have at least the declared coverage (asymptotically) regardless of the uncongeniality. And this bound is achievable with extreme uncongeniality, that is, the possible over-coverage due to doubling variance is a very reasonable premium we pay to insure against any degree of uncongeniality.

# 3 Principled Corner Cutting: A Car-Talk Bayes Theorem

## 3.1 A Car Talk puzzler

On July 11, 2015, I was driving to my dentist's office when I heard on the Car Talk radio broadcast the answer to its previous week's puzzler (July 4th). One of the host brothers, Ray Magliozzi, was rendering, in a slightly naughtier way, the text version in `https://www.cartalk.com/content/possible-false-positive`, reproduced below.

> RAY: There's a rare disease that's sweeping through your town. Of all the people who are exposed to it, 0.1 percent of the people actually contract the disease. There are no symptoms until the disease actually occurs. However, there's a diagnostic test that can detect the presence of the disease up to a year before it strikes.
>
> You go to your doctor, and he administers the test. It comes out positive. You say, "I'm done for!"

> Then you get a little bit encouraged. You say, "Wait a minute, doc, is this test 100 percent accurate?" Your doctor responds, "Well, not really. It's 95 percent accurate." In other words, 5 percent of the people who take the test will test positive but they don't really have the disease.
>
> Here's the question: What are the chances that you actually have the disease?

My heart pounded faster, literally. I was excited. Another great story about Bayes theorem that I can use for public lectures and introductory courses—I am always on the lookout for materials that can help to engage statistically innocent audiences. I was anxious. What would be Ray's answer—would he get it right? I was worried. How could he get it right—there is not enough information to apply the Bayes formula!

I had to hold my breath (and the steering wheel) when Ray gave his answer. Again, the following text posted on Car Talk website is a bit crisper than the radio version.

> RAY: Let's say 1000 people take the test. Fifty people will test positive and yet they will not have it. One will test positive and have it. So your chances of actually having it, even though you tested positive, are one in 51, or a little less than 2 percent. So who's our winner?
>
> TOM: The winner is – wow! Frank Migliozzi from Rye Brook, New York! Congratulations.

WOW indeed, but not for keeping the winner almost in family. Ray had accomplished something that a card-carrying statistician like myself would never try, at least not publicly. That is, invoking an incorrect or at least incomplete reasoning to handle incomplete information to reach an approximately correct answer—letting two wrongs somehow cancel each other.

Ray's reasoning was clearly incomplete because the Bayes theorem requires three rates (prevalence, specificity and sensitivity), but the original problem assumes only two rates (1% and 95%). Indeed, it seems that whoever originally posted the problem (which may or may not be Ray) may not even realize the concept of false negative, because the problem appears to interpret "95 percent accurate" as *specificity*, that is, the percentage of people in the healthy population who will be declared as such by the test (but see Section 3.4 for an alternative to this interpretation). Hence the problem only specified false positive rate, which is one minus specificity. The false negative rate, which is one minus *sensitivity*, can be anything. So how could there be a unique answer, or any meaningful answer?

## 3.2 Single-error bounds on Bayes risks for rare events

Intrigued by Ray's argument, I decided to get to the bottom of it, an exercise which also helped to distract me from my or(al)deal. Let $p$ be the prevalence rate of a specified disease in a given population, and $f_-$ and $f_+$ be respectively the false negative and false positive rates of a chosen test for that disease to be administered in the population. Then Bayes theorem tells us that the probability of a *randomly* selected individual with a positive test result who actually suffers from the disease—the so called positive predictive value—is

$$B = \frac{p(1-f_-)}{p(1-f_-)+(1-p)f_+} = \frac{p}{p+\eta f_+}, \tag{3.1}$$

where $\eta = (1-p)/(1-f_-)$, which is well defined as long as $f_- < 1$. (Let's hope no one uses a test with 100% false negative rate.) Formular (3.1) immediately suggests that, as long as $p \le f_-$, that is, $\eta \ge 1$, then

$$B \le \frac{p}{p+f_+}. \tag{3.2}$$

But the right-hand side is exactly the expression Ray used! Therefore, Ray's method actually delivers a safe upper bound for assessing risk for rare events, as long as the rarity does not exceed the false negative rate, which typically is a safe assumption to make. Indeed, it is easy to see from the above derivation that the inequality (3.2) holds if and only if $p \leq f_-$.

In biostatistics and epidemiology, a handy rule of thumb has been that for a rare disease and a reasonably good test, specificity matters far more than sensitivity in determining the positive predictive value; see for example Van Belle et al. (2004, *p.* 559) or Blitzstein and Hwang (2019, *pp.* 43-44). The inequality given in (3.2) rigorously establishes this fact, telling us exactly when and how to use this rule as a safe upper (or lower) bound, and explicating the meaning of "reasonably good" (i.e,, $f_- \leq p$).

Somewhat unexpectedly, however, the elegant if-and-only-if result also holds with the overall error rate:

$$f_o = \Pr(\text{the test renders a wrong diagnosis}) = pf_- + (1-p)f_+. \tag{3.3}$$

That is, we have the following intriguing result

$$B \leq \frac{p}{p+f_o}, \quad \text{if and only if } p \leq f_o. \tag{3.4}$$

This result says that, should Ray's 5% error be the total error instead of false positive rate, his numerical answer would still provide a rather useful upper bound. These bounds are useful because they are rather close to the actual $B$. For example, assuming $f_- = f_+ = 0.05$, then $B = 1.87\%$, and the bound in (3.2) (which is the same as (3.4)) is 1.96%. Such a difference between them is inconsequential for many practical purposes, yet the upper bound is considerably easier to obtain instantly: adding up prevalence and the total error rate, and divide the prevalence by this sum. Such a practical tool can help to enhance general professionals' (e.g., doctors, lawyers) as well public's ability to obtain reasonable risk assessments, and hence to increase their chances to spot flawed reasoning and false conclusions. For example, no prosecutors can convince a jury by arguing 2+3=6, yet they can easily impress on uninformed juries by declaring "Look, there should be little doubt that the defendant is guilty because the blood test is 99% accurate." Those who have listened to Ray's reasoning would at least have a chance to remind themselves: "'Wait a minute ...".

Given the potential use of such single error rate bounds and approximations, Theorem 1 below collects three bounds and provides the error assessments when these bounds are used as approximations to $B$ of (3.1).

**Theorem 1 (A Dirtified Bayes Theorem)** *Let $D$ be an event for which we wish to assess risk, with prior risk $p = \Pr(D)$. Let $T$ be a test for assessing if $D$ is present, with $f_- = \Pr(T = -|D)$ and $f_+ = \Pr(T = +|D^c)$ being its respectively false negative and false positive rates, and with $f_o = pf_- + (1-p)f_+$ being its overall error rate. Then for assessing the risk of $D$ after a positive assessment $T$, $B = \Pr(D|T = +)$, we have the following three sets of results.*

**(I) When only $f_o$ is available.** *Let $O_- = f_-/(1-f_-)$ be the odds of committing false negative and*

$$B_o = \frac{p}{p+f_o}, \quad \text{then} \quad \delta_o \equiv \frac{B_o - B}{B} = \frac{f_o - p}{p+f_o} O_-. \tag{3.5}$$

*In particular, $B \leq B_o$ if and only if $p \leq f_o$, and also $B = 1/2$ if and only if $f_o = p$. Furthermore,*

$$|\delta_o| \leq O_-, \quad \text{for all } p \in [0,1]. \tag{3.6}$$

**(II) When only $f_+$ is available.** *Let*

$$B_+ = \frac{p}{p+f_+}, \quad \text{then} \quad \delta_+ \equiv \frac{B_+ - B}{B} = \frac{1-B_+}{1-f_-}(f_- - p). \tag{3.7}$$

*In particular, $B \le B_+$ if and only if $p \le f_-$. Furthermore*

$$|\delta_+| \le \max\left\{O_-, \frac{f_+}{1+f_+}\right\}, \quad \text{for all } p \in [0,1]. \tag{3.8}$$

**(III) When only $f_-$ is available.** *Let $O_p = p/(1-p)$ be the prior odds for $D$ and*

$$B_- = \frac{O_p}{O_p + O_-}, \quad \text{then} \quad \delta_- = \frac{B_- - B}{B} = \frac{(f_+ - f_-)}{(1-f_-)O_p + f_+}. \tag{3.9}$$

*In particularly $B \le B_-$ if and only if $f_- \le f_+$. Furthermore, we have*

$$|\delta_-| \le \frac{|f_+ - f_-|}{f_+}, \quad \text{for all } p \in [0,1]. \tag{3.10}$$

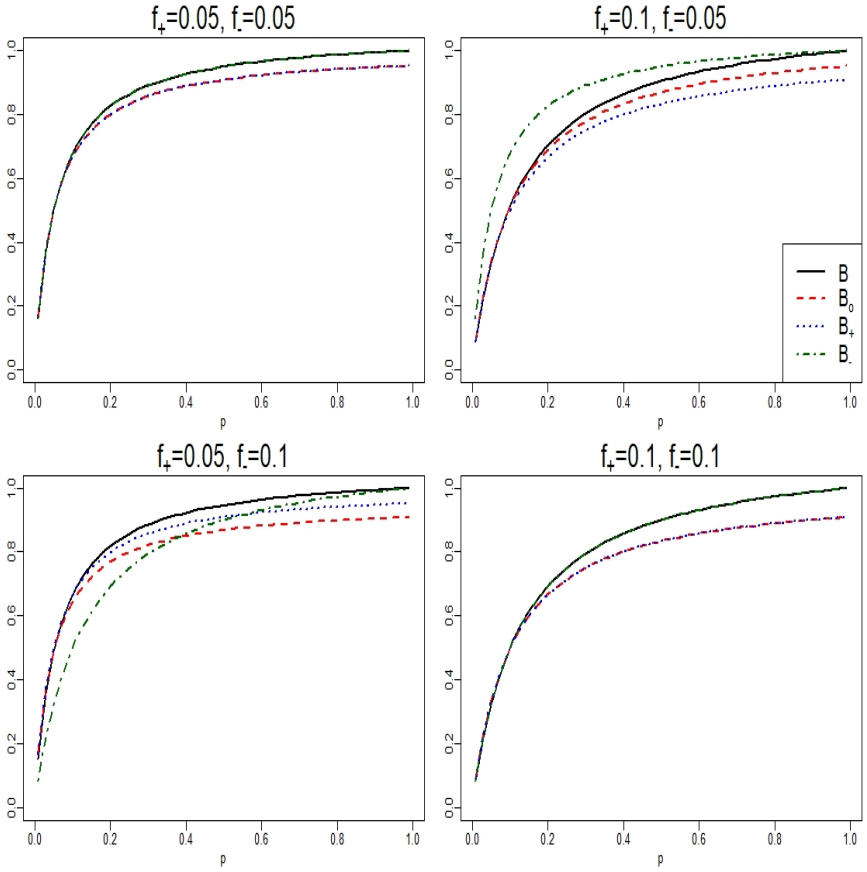



Figure 1: Demonstrating the absolute errors in approximating the posterior probability $B$ by three bounds: $B_o$ (requiring overall error rate $f_o$; dash lines), $B_+$ (requiring false positive rate $f_+$; dot lines), and $B_-$ (requiring false negative rate $f_-$; dot-dash lines). Four cases by crossing $f_+ \in \{5\%, 10\%\}$ with $f_- \in \{5\%, 10\%\}$

Part (III) deviates from the previous two bounds because it operates with odds (slightly harder to calculate mentally), and the condition on being a bound is not driven by the rarity of the event, but by comparing the false positive and false negative rates. It is included here for completeness with cases where one has access only to false negative rate $f_-$ but still needs to roughly assess positive predictive value $B$. Because $B$ is largely determined by $f_+$ as discussed earlier, the only way that we can still have a reasonable bound or approximation is when we possess some knowledge about the proximity between $f_+$ and $f_-$. It therefore should not come as a surprise that in general the bound in (III) can work badly (or really well) regardless the magnitude of $p$. Figure 1 (absolute error) and Figure 2 (relative error) illustrate this point via the plots with $f_- \neq f_+$. They also illustrate that the first two bounds in Parts (I) and (II) work rather well as approximations, with relative errors in approximation not exceeding 11% across board, as anticipated by (3.6) and (3.8) (in this case, both relative errors are controlled by $O_- \leq 1/9 \approx 11\%$).

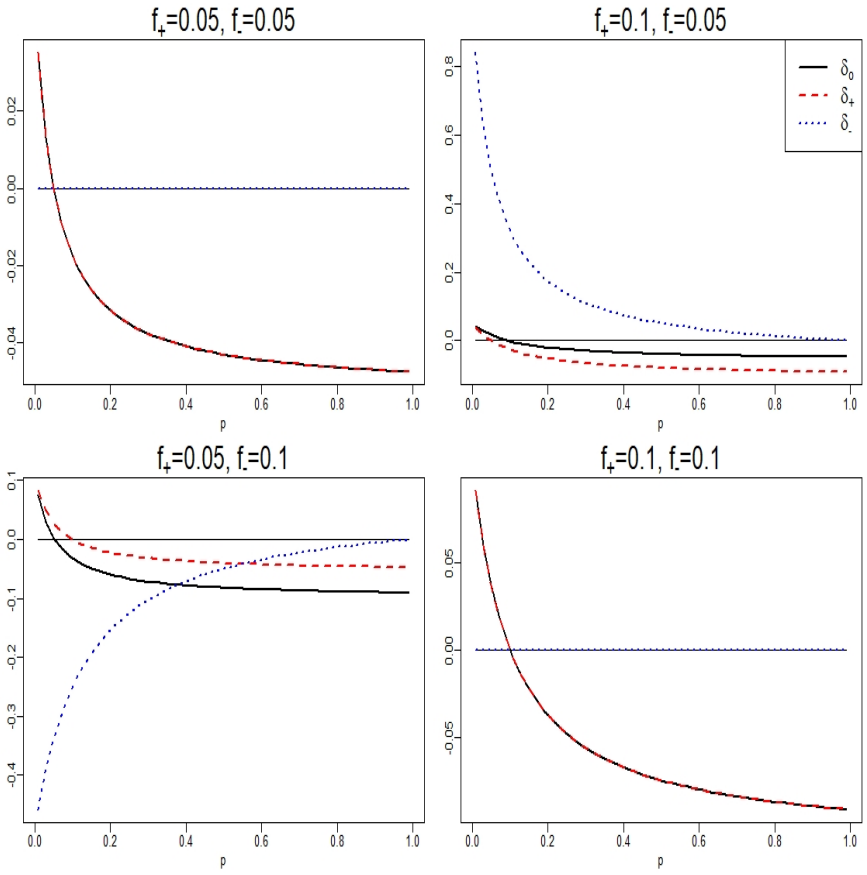



Figure 2: Demonstrating the relative errors in approximating the posterior probability $B$ by three bounds: $B_o$ (requiring overall error rate $f_o$; dash lines), $B_+$ (requiring false positive rate $f_+$; dot lines), and $B_-$ (requiring false negative rate $f_-$; dot-dash lines). Four cases by crossing $f_+ \in \{5\%, 10\%\}$ with $f_- \in \{5\%, 10\%\}$

## 3.3 Principled Corner Cutting (PC2)

The dirtified Bayes theorem revealed above is a very effective illustration of what I have been advocating: *principled corner cutting* (PC2) (see `http://videolectures.net/nips2010_meng_mlhi/`). In the grand scheme of things, any scientific study must cut corners both for its feasibility and utility. For example,

applying Bayes theorem would neither be feasible nor useful if we were to insist on modeling all unknowns (jointly) given all knowns. In general, the moment we adopt an assumption, we effectively cut out all the considerations that contradict the assumption. And we inevitably cut corners due to constraints from all directions: data limitations, computational challenges, knowledge inadequacies, time and funding shortages, etc. Therefore, the differences among good, bad, and ugly studies lie mainly in what corners are cut, according to what principles, resulting in what consequences, and is it worthy to restore them? That is, PC2 has three main parts:

**(A)** Guide and prioritize the corner cutting with theoretical understanding and pragmatic acumen under practical constraints;

**(B)** Identify and cut corners with controlled consequences and approximate constrained optimality;

**(C)** Understand if it is wise or even possible to seek additional resources and information for restoring the corners cut, and if so, in what sequence and how to restore.

For Ray's problem above, the guiding principle needed by (A) is probabilistic assessment via Bayes theorem, and the practical constraints are (i) lack of full information for applying Bayes theorem, and (ii) the method and reasoning need to be accessible to the general public. The corner cutting in (B) is to replace the exact answer, which is unknowable without further information, by an upper bound, which is far easier to calculate and explain than the full Bayes theorem. Indeed, although Bayes theorem has been around for centuries and has entered popular media[5], explaining Bayes theorem and its ingredients (prevalence, specificity, and sensitivity) to the general public is still a demanding task, as demonstrated in the grand finale article by Waller and Levi (2021) for the COVID-19 issue in *HDSR*.

For (C), this is the type of situation where insisting on getting more information (e.g., the false negative rate) in order to obtain the exact principled answer (e.g., from Bayes theorem) would make us statisticians irrelevant in the eyes of practitioners, because they simply do not have that kind of luxury. But this does not make the principle itself—in this case Bayes theorem—irrelevant. To the complete contrary, it is precisely the understanding of Bayes theorem that renders us the insight on when and why Ray's argument works, and how well it does compared to the ideal answer.

This last comparison is particularly important for answering the practical question: "Is the answer good enough for my purposes?" In this case, the answer is almost surely a yes. It is rare for a medical screen test to have a false negative rate less than 0.1%, and the difference between 1.87% or 1.96% is inconsequential for either patients or doctors in their decision making. The more refined answer, even if it were available, would have essentially zero practical impact in such cases.

Whereas the benefits of PC2 are obvious from both practical and economical perspectives (e.g., in terms of human capital investment for research), its routine adoption is far from trivial. Practicing PC2 requires minimally good understanding of pros and cons of the available tools, and their applicability and cost of implementation in a particular context. Good judgments under time and other constraints are key for its success and for achieving its maximal benefits. Neither of them can be learned effectively in classrooms or from textbooks alone. To make the matter worse, currently, we have far too few educators with sufficient PC2 experiences to design and teach PC2-oriented curriculum and training programs. But good progress is being made, especially as a part of the broader effort to ensure statistical education meets the general demand

[5]See `https://www.newyorker.com/books/page-turner/what-nate-silver-gets-wrong` or `https://www.nytimes.com/2020/08/04/science/coronavirus-bayes-statistics-math.html`

of data science, where acquiring pragmatic acumen is being recognized as a central skill for successful data scientists; see a series of articles in *HDSR*, e.g., Haas et al. (2019); Berthold (2019); Fayyad and Hamutcu (2020), and especially Kolaczyk et al. (2021) and its eight discussions (followed by a rejoinder).

### 3.4 Avoid *head waving*: Cut corners, not principles

Further discussions on how to improve the general education on statistical thinking and insight, a critical competency for executing PC2, are presented in Section 5. Inspired by a previewer's comments, here we discuss a closely related competency issue: the ability of fluently transitioning and translating between the cognitive language for intuitive thinking and probabilistic language for formal reasoning. This issue is particularly tricky and consequential when we reason with conditional probabilities. Considering Ray's interpretation of the "95% accuracy" in the Car Talk example: "In other words, 5 percent of the people who take the test will test positive but they don't really have the disease." Using our notation, does Ray's descriptive language capture the conditional $\Pr(T = +|D^c)$ or the joint $\Pr(T = +, D^c)$? Given the description is about a group of people who share two characteristics: test positive and disease free, interpreting it as a joint probability seems more logical. Indeed, if one argues that this description is about a conditional probability, then it is more a description for $\Pr(D^c|T = +)$ than for $\Pr(T = +|D^c)$ because the first stated condition is the test being positive.

Those who have training in probabilistic languages would be careful to avoid such ambiguous descriptions. The issue here is not merely communicating clearly, but rather that the descriptive language reflects our thinking, especially its ambiguity, which can seriously mislead us. As a previewer pointed out, if we interpret Ray's descriptions as about joint probabilities, then his statement that "Fifty people will test positive and yet they will not have it" (out of 1,000) specifies $\Pr(T = +, D^c) = f_+(1 - p)$, and "One will test positive and have it" gives us $\Pr(T = +, D) = (1 - f_-)p$. Ray's formula using the second quantity divided by the sum of the two then would yield $\Pr(T = +, D)/\Pr(T = +)$, which is exactly the positive predictive value $\Pr(D|T = +)$! In other words, we can either take Ray's reasoning as an approximation leading to an upper bound as established in the dirtified Bayes Theorem, or conclude that it is Ray's ambiguous descriptive language that injects assumptions that were not given by the original question, which made it possible for him to arrive at a complete and unique answer.

This is a case of "gain in translation", which is no less troublesome than lost in translation. As a matter of the fact, many inference puzzles, such the prison dilemma and Simpson's paradox, are consequences of injecting assumptions that are not given, and sometimes even in logically inconsistent ways; see for example Gong and Meng (2021) for an overview and delineation. An effective way of identifying and avoiding influentially consequential ambiguity is to put intuitive contemplation in writing using mathematical notations with explicit meanings. Mathematical languages are most challenging precisely when they are most in need, because they exposes the laziness in our "head waving", such as bypassing the step for explicating what we mean or without examining internal inconsistencies.

A key message here is that for PC2 to work well, we must not cut corners on principles, whether for theoretical insights or for practical acumen (Cox, 2006; Cox and Donnelly, 2011; Reid and Cox, 2015; Hacking, 2016; Cox and Snell, 2018). Indeed PC2 demands higher and deeper levels of principled contemplation for foreseeing consequences more steps ahead. Perhaps a reasonable analogy is that master chess of Weiqi (i.e., Go) players have better ability than average players to see farther ahead the consequences of each move, an ability that permits them to make seemingly foolish sacrifices to spectators, and yet decisively winning

strategies. Statistical and data analytical thinking is even harder because the rules of the (soft elimination) game are less clearly laid out. Indeed, without sufficient rules and principles to tame our innate desire for success, head waving often gets us into trouble because of its cherry-picking tendency, fueling self-fulfilling prophecies rather than self-critical introspection, the topic of the next section.

# 4 Quality Introspection: The Pufferfish/Selfish Test

## 4.1 Thou shalt not sell what thou refuseth to buy

Pufferfish, known as *Fugu* in Japan and *Hetun* in China, is a delicacy. That is, if it does not kill you. In Japan, being a Fugu chef requires a license. Legend[6] has it that to obtain a license requires a 2-3 year apprenticeship with a licensed Fugu chef, and then an examination. This exam must be among the world's most well prepared ones, because it involves a very practical test: eat what you prepare. This surely increases consumers' confidence in being served by Fugu chefs—they are alive.

(a) A Northern Pufferfish caught in Long Island's Great South Bay, and was released back into the water alive and well.
(Photo by Brian Yurasits on Unsplash)

(b) An alleged *Hetun* for a Shanghai lunch, and its diner was released back into the city alive and well, but unimpressed.
(Photo by an involuntary diner on alert.)

Figure 3: The only reason that I was willing to let my palate do the thinking was because the chef had done the same (but this particular chef might have used $p < 0.005$ instead of $p < 0.05$, because the result was too safe to be a delicacy; rumor has it that the deliciousness of a *Hetun* comes from a carefully calculated dosage of its poison.)

I doubt that there is any statistical delicacy worth dying for. But we can institute a similar *selfish* test. If I am ready to write about my wonderful data analysis to show that a new treatment is the best for a serious disease, then surely I'd request that treatment for myself or a loved one, if (God forbid) I or my loved one contracts the disease, right? Similarly, if I have shown how a new education program is at least twice as effective as any existing ones, then (of course) I'd place my kids into that program, correct?

If any hesitation arises in answering such self-questioning, then we owe it to ourselves and our profession a pause and some introspection. Without hesitation does not necessarily imply high confidence, since each of us has a different tolerance for risk, but at least we do not impose on others the risks that we are not willing to take ourselves. We know best what we have done or not done, the judgments rendered or self-overruled,

[6] See `https://www.nytimes.com/1981/11/29/travel/one-man-s-fugu-is-another-s-poison.html`

the criticisms accepted or rejected, the shortcut made that should not have been taken, the incentives for rushing that should have been resisted, etc. We may never tell anyone about all the defects for which we would take points away if they had appeared in our students' projects. But our professional consciousness—if we have one—should remind ourselves of how we made our sausages, when we consider ourselves consumers of our own products. If there is anything we find hard to swallow, then we should not serve it to others, or at least not without serious warning. This is in the same spirit as the mantra in the business world: "Eat your own dog food", though the assertion[7] that "If a dog food is of the high quality advertised to consumers, then it should be good enough for a person to eat as well" itself is an ironic demonstration of lack of introspection: how do humans know that our pecking order for the food quality is shared by dogs?[8]

Among all professional ethical considerations, practically motivated or ideologically driven, the mantra "Don't sell what you refuse to buy" should constitute the most basic professional benchmark for decency, just as "Don't treat others the ways you don't want be treated" reflects the golden rule for human decency. Indeed, this personal introspection can be viewed as an attempt to formulate and achieve empathy-driven objectivity, complementing the notion of scientific disinterestedness (in parallel to aesthetic disinterestedness, see for example Came, 2009), which share the same goal of ensuring scientific reliability but by removing one's interests instead of injecting them.

Ideally, our introspection should take a critical look at the entire process that produced the results we want to scrutinize. In reality, this step itself will suffer from various omissions—most of us cannot work more than 12 hours a day without damaging our health or relationships. Therefore, the checklist provided below should be considered as an introspective menu for us to look through and to choose from, depending on which aspects of our process are more likely to make our results fishy, just as we tend to choose the most savory items from a restaurant menu.

- Did I understand and consider carefully the data collection and pre-processing processes in my study? How much do I know about the quality of the data I used?
- Did I have sufficient understanding of the substantive problem to recognize its inherent challenges, such as confounding factors and lack of identifying information?
- Are the assumptions and models I adopted free of internal contradictions? If not, what justifications do I have to allow such contradictions?
- Did I follow principled and holistic methods such as probabilistic propagation and conditioning or did I rely on ad hoc "intuitive" methods?
- What approximations did I make in modeling, mathematical derivations, or computation? Which ones are most vulnerable?
- Did I understand the impact of these approximations, especially the worst damage they can cause? Have I looked at studies on their impact, or have I investigated them theoretically or empirically?
- Did I perform sufficient validation or robustness check for models and assumptions I posited? Am I confident that my results would hold reasonably well if my data are perturbed somewhat?

[7] See, for example, `https://www.investopedia.com/terms/e/eatyourowndogfood.asp`

[8] Presumably this can be studied via a blind testing, in the fashion of Judgement of Paris (see `https://www.cnn.com/travel/article/judgment-of-paris-wine-tasting-cmd/index.html`), which was very effective in elevating underdogs; some clever design, however, is needed to record dogs' preferences.

- Did I carry out reliability checks on the numerical evaluations or simulation, such as verifying computational results independently by two different methods, routines, or even research assistants?
- Did I commit any form of cherry picking, from data gathering to results validation? If so, what reasons do I have to believe that my cherry picking would not do much damage?
- Do I understand the findings at a level that would enable me to explain and teach them confidently to non-experts?

I am sure many readers will find this list incomplete or even inappropriate. The emphasis here is not on any particular way of conducting the introspection, but on having it as an integrated part of our quality control before we sign off on any study for which we serve in our professional capacity. The more we institute such self-scrutiny, the better we serve science, society, and our profession. Yes, it is not an easy process to do well, and it is not rewarded or even recognized by many of the current incentive systems. But by putting ourselves in the shoes of those whose lives or livelihoods will be affected by our analyses, we can be much more mindful in making assumptions, choosing methods, cutting corners, interpreting results, etc. Delivering what we promise is one way to gain public trust. Being self-critical is another, especially when we can establish it as a professional culture, just as particle physicists have (see Junk and Lyons, 2020). Furthermore, by routinely engaging in such a practice, we also encourage anyone who conducts statistical or data analytical investigations to do that same. We statisticians—myself included—are often frustrated from seeing abusive or even just shallow statistical analyses, and a part of the frustration is that even if our entirely profession gives up sleep, we still would not have remotely sufficient human power to eliminate statistical nightmares, so to speak. Our best bet is to lead by example, help to incentivize quality control, and inoculate future generations with a mindset for appreciating the world of uncertainty and the uncertainty world.

## 4.2 Incentivizing quality introspection

Introspection via selfish test relies on internal reward systems such as peace of mind or a sense of professional pride, because our (any?) profession has no effective ways to enforce it. But it is possible to conceive external incentive systems that can encourage higher levels of self-scrutiny, if we allow ourselves to contemplate the notion of *behavioral statistics*, to paraphrase *behavioral economics* (e.g. Mullainathan and Thaler, 2000; Camerer and Loewenstein, 2004; Wilkinson and Klaes, 2017), a rather self-explanatory term.

To start, consider incentives for publication in universities regarding appointments, tenure, salary raises, etc. During my deanship (2012-2017), I surprised myself for having succumbed to becoming a bean counter, despite my best efforts to avoid it. It is an NP-hard (Not Practical) problem for a dean to read even just a single article from each candidate to form a direct sense of research quality instead being impressed (and imposed upon) by candidates' CVs and others' testimonies of their accomplishments. Even if I were given all the time to do so, I'd not have the basic knowledge to understand the key messages of most articles outside of my knowledge stream, which sadly is not even an epsilon compared to the ocean of knowledge that a dean is effectively asked to navigate. Consequently, I constantly caught myself counting the number of articles and books or the number of awards in a CV, and I was not alone in dealing with such a reality. Granted, promotion and hiring decisions should be and are made collectively, but each of us should provide our opinions informed by at least some understanding of quality of the candidate's work instead of quantities. The experience reminded me of a broader lesson: any mechanism for discouraging trading quantity for quality

is most forceful when it is incentivized to be self-enforced before the product is made. Relying on external enforcement is often too late and weak, because realities can easily turn our best wishes into wishful thinking.

Understanding this reality, a brave university may announce that it would (permanently) deduct $\alpha$% of a professor's salary if some of the professor's published statistically significant results at the $\alpha$ level turn out to be wrong. A 5% salary reduction is non-trivial for most academics, especially considering its compounding effect. In contrast, a 0.5% reduction is much more tolerable, even with its compounding effect, considering the length of a professor's serving years. Granted, proving that any statistically significant result is wrong is a daunting (and unpleasant) task, and using a smaller $\alpha$ may delay or even eliminate a publication. But why risk it if it is under the control of my choice of the $\alpha$ level? I still can choose $\alpha = 5\%$ if I'm so sure of the results and hence willing to take the risk. But this is exactly the thinking process such incentive systems aim to encourage: the more researchers are incentivized to self-control quality, the fewer non-reliable studies would leak into scientific literature.

"Xiao-Li, you are hallucinating—no university would ever consider such a laughably naïve and frankly dangerous idea!". Very true. Such a system can penalize productive faculty (though it is partly intended to discourage chasing quantity instead of quality), hurt collaborations (why should I be penalized for my collaborators' sloppiness over which I have no control?), and even stifle creativity (in fear of making costly mistakes).

However, it might not be a completely crazy or harmful idea for regulatory agencies, such as the FDA, to consider incentives that would encourage applicants' self quality controls beyond established standard requirements. For example, those applicants who volunteer to impose a more stringent criterion (e.g., a smaller $\alpha$ level) would be given accordingly more benefit of doubt, such as a higher bar for requiring them to withdraw drugs from markets when post-approval complications arise. Or the priorities in the approval process depend on the degree of self-quality control, e.g., the smaller the $\alpha$ the higher priority. The exact scheme is less important than introducing a *quality control knob* that allows the applicants to dial to optimize over their own risk and economic considerations. As long as the dial is set to minimally maintain the current standard (e.g., $\alpha$ must not exceed 5%), the self-incentivized system can only improve upon the current practice.

Of course, realities typically are more complicated than what we conceive. For example, an adjustable incentive system can and will induce more serious gaming behaviors or even fraudulent manipulations. But these complications are expected in any system, and we can deal with them as a part of many trade-off considerations. For instance, in the drug approval context, Chaudhuri et al. (2020) considered the trade-off permitting a larger $\alpha$ in exchange for a shorter clinical trial period involving anti-infective therapeutics during pandemic outbreaks, while minimizing the expected harm of false positives and false negatives. Their findings of $\alpha$'s being as large as 26% would be considered too radical to be entertained during a normal time. But during a pandemic, they could be the optimal numbers, saving more lives (and livelihoods) than otherwise. It is in the same spirit that we should permit ourselves to explore more radical incentive systems as a part of our effort to reduce irreplicable or more critically unreliable studies (Meng, 2020a).

### 4.3 Incentivizing more *behavioral statistics*

Methodologically, instituting any of such incentive systems would also encourage and challenge ourselves to systematically study *behavioral statistics*, that is, statistical modeling and analyses that inherently take into account the behavior of entities involved in the study. This is in parallel to *behavioral economics*, which arose

because of the general recognition that the traditional "rational choice" framework (Becker, 1976), although mathematically convenient, is too ideal to capture how individuals and organizations behave in reality (see, e.g. Samson, 2016). In this sense, behavioral statistics is not new at all, and in fact one can argue that many statistical concepts and methods cannot be rigorous or applicable without being *behavioral*.

For example, it is well known (see e.g., Little and Rubin, 2019) that to sensibly handle non-responses in surveys, we must take into account the responding behaviors of the surveyed individuals. Assuming that people response randomly, the so-called missing completely at random (MCAR) model (Rubin, 1976), would be extremely convenient in theory and for computation, because then the observed sample inherits all the good (and bad) properties of the surveyed sample, but just with a smaller size. In reality, however, MCAR is extremely rare. Worse, a seemingly small deviation from MCAR can easily destroy any confidence we can place in the survey estimates if we fail to correct for the non-MCAR behavior; a most striking recent example is for predicting US presidential elections (Meng, 2018). The realization of the importance of such behaviors has led to a large literature on studying missing data mechanisms (Rubin, 1976; Heitjan and Rubin, 1991), which is the hardest problem to deal with among the three broad class of complications created by missing data or more broadly by incomplete data (Meng, 2012).

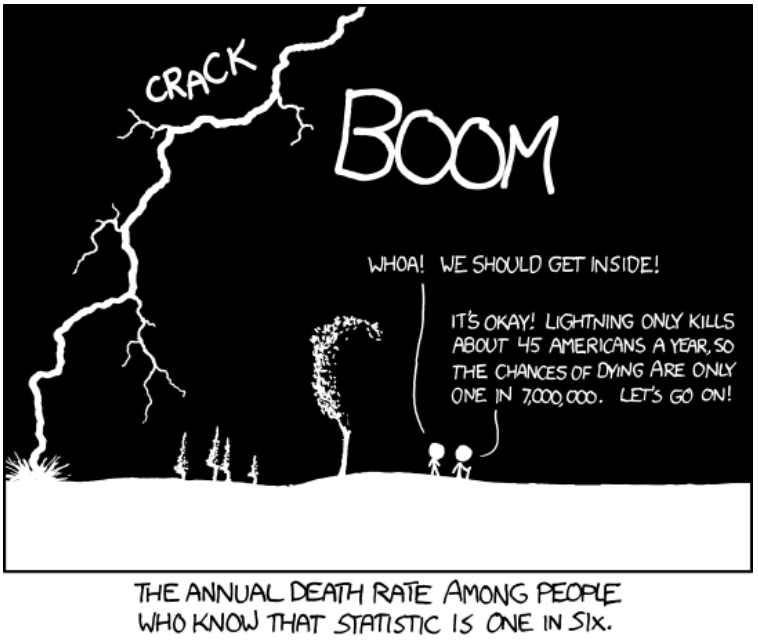



Figure 4: An Enlightening Illustration of *Behavioral Statistics*, created by `https://xkcd.com/795/`. Reproduced under the CC BY NC 2.5 Licence as granted by `https://xkcd.com/license.html`. [`xkcd` has produced a host of thought-&-smile provoking cartoons, and this is just one of many.]

To encourage ourselves to capture as much reality as possible, it is useful to contemplate questions such as: What will be the actual false positive rate among those studies that choose to adopt $\alpha = 0.005$? Would it be higher or lower than 0.5%, and what are the determining factors and conditions (e.g., on researchers' ability to assess risk) in driving the directions? Whereas such questions are not easy to answer, the following `xkcd` cartoon vividly reminds us of the importance to at least to contemplate their potential impact — at the end of the day, what really matters is not governed by our wishes or idealization but how we—and our procedures and products—behave in reality. Such contemplation can also push us to revisit our theory and principles,

an upward iterative process that is a hallmark of the scientific progress. For example, is *significance level* even a meaningful concept when every study is incentivized to choose its own? This question is no longer hypothetical when we follow the proposal to "justify your alpha" (Lakens et al., 2018), which makes a rather sensible (and obvious) point that a universal $\alpha$, regardless of its value, is undesirable. But when everyone does justify their own $\alpha$, then a "conditional risk" question highlighted by the `xkcd` cartoon becomes a very relevant one for regulatory agencies such as the FDA at both a conceptual and practical levels.

# 5 Kidstogram: Let's Plant Some Random Seeds

## 5.1 A big hole in our elementary education

Among reasons for unreliable statistical and scientific studies, from innocent mistakes to skillful deceptions, we tend to overlook perhaps the most evasive and invasive one: a big hole in our elementary education for seeding an appreciation of the necessity and beauty of uncertainty. From the moment we teach our children counting by fingers, we embark on a mathematical journey designed for developing brains to navigate multiple mazes of rules and formulas for understanding and manipulating deterministic relationships. We seldom give them a tour of a random forest, or even show them an ambiguous path to a stochastic land. Yes, even when we teach statistics, we tend to teach it as a set of rules and in the order of (mathematical) complexity: one sample test, two sample tests with equal variance, two sample tests with unequal variance, etc. We teach linear regression as line fitting, treating uncertainty as annoying "residual errors" to be gotten rid off, wasting great opportunities to intellectually inspire and enhance the young minds (Meng, 2009a, 2010, 2009b).

I venture to argue that the very reason that most of us—myself included—feel uncomfortable in dealing with uncertainties, whether in life or in work, is that our pre-college education system has failed to accustom our brains to appreciate uncertainty as much as information. We have failed, collectively, to teach that uncertainty and information are the two sides of the same coin: *variations* (Meng, 2020b). For example, if everyone at a train station has identical appearance, then any description of an individual there would contain zero information for identifying the individual. We would be as annoyed (and frightened) as if we were given no description. When taught earlier and given sufficient exposure and opportunities, we can all appreciate uncertainties just as we desire for information. Statistically speaking, there should *not* be any particular group of brains that are more suited for processing uncertainty than others, just as regardless of the race or ethnicity of a group of children, raise them in any language, and that language will become their mother tongue.

On the other hand, it is much harder to acquire a second language, and the difficulties generally increase with the starting age. I have been living in an English speaking environment for 35 years as of this writing, and I am still struggling with every article or speech, from basic grammar to common pronunciation. I simply do not have the innate and instant feeling when to use "a" or "the", for instance, and have to make a particular effort to avoiding mixing "she" with "he" in pronouncing since this phonetic distinction does not exist in Chinese. Thinking analogously, I am fully sympathetic to all deterministically trained minds struggling with stochastic realities. Take linear regression as an example. When we teach it as fitting a line, how confusing it must be that predicting the sale price of a condo from its rent cannot be read off from the same line as for predicting the rent from its sale price?

And indeed linear regression is a great example to highlight the inadequacy of our current educational preparation for understanding stochastic relationships. Even for those who feel comfortable dealing with

uncertainties in their estimates, many of them fail to emphasize or appreciate that the most important reason for assessing uncertainty appropriately does not concern the error bars. Much more critically, it is about properly propagating uncertainties for our inference and prediction, because the uncertainty may change our estimators in fundamental ways. This is the very reason that converting $y = \beta x$ into $x = y/\beta$ would lead to a mathematically provably inferior prediction for $x$ from $y$ (under square loss), even if this conversion seems to be the only sensible algebraic rule.

Collectively, we have been breathtakingly innovative in training generations of young brains with ability to manipulate deterministic rules in the most effective way. I do not use the phrase "breathtakingly" lightly—my breath was taken away when I was watching a "hand waving" method[9] used during an arithmetic competition. For those of us who do not understand how this method works, it is a bizarre and fascinating scene. A group of elementary school students wave their left hands rapidly and in seemingly chaotic ways, yet simultaneously their right hands are writing down answers continuously as their eyes are scanning through over 220 arithmetic problems in 15 minutes, with a reported success of reaching 219 correct answers. In a similar vein, for those who have no exposure to statistical thinking, it must be equally bizarre when I explain to them that $y = \beta x$ does not imply $x = y/\beta$, and fascinating that I, apparently algebraically challenged, can actually navigate the protean world of data.

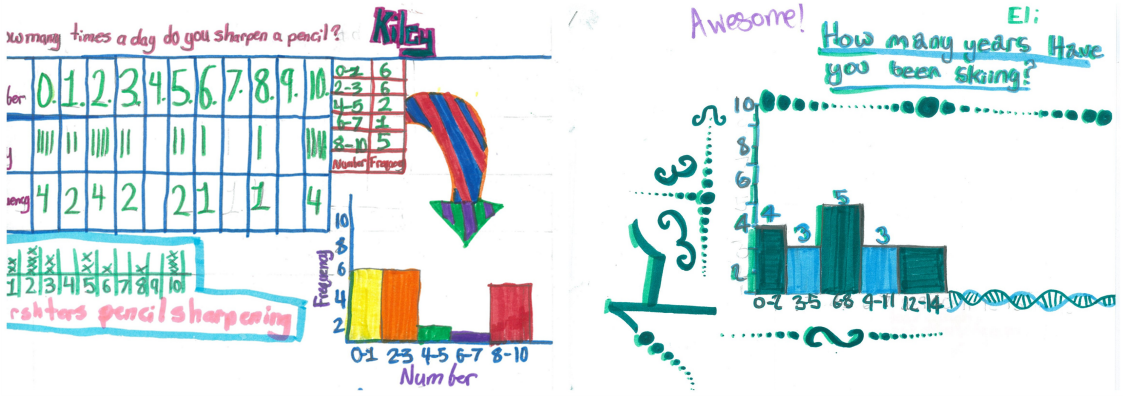
If we can be so innovative for teaching a subject as old as arithmetic, then surely we can put our creative minds together to thrust ahead a path for early childhood education that is based on appreciating and internalizing the concepts and vocabularies of variations and uncertainties. The sooner we receive such training, the more fluent we all become in speaking the language of variation later in life. Indeed, teaching histograms can start immediately after learning about counting, since a histogram is nothing but an ordered bookkeeping of counts.

## 5.2 Seeding distributional thinking in early childhood education

Probabilistic calculations and manipulations are challenging for many of us because they do not operate with numbers, but rather with distributions. The advance from a single count, i.e., a number, to an ordered collection of counts, that is, a histogram, is far more epistemological than mathematical, because histograms compel *distributional thinking*. As Sanders (2020) summarized nicely, "distributional thinking can be defined as the frame of mind for considering the outcome of a process as not just a singular state of being, but rather a pattern of alternatives and their likelihoods." Some might wonder if it is possible at all for a developing brain to comprehend patterns and processes when it still struggles with counts and rules. I'd argue that it is not only possible but actually it should be easier to engage developing brains in distributional thinking because patterns are more pictorial than numbers, and processes are more participatory than rules.

Just as a proof-of-concept illustration, Figure 5 showcases two histograms by children from the *One Room Schoolhouse,* an innovative lab school in Denver, Colorado (Burt, 2014). Or we should really term them as *kidstograms*, not merely because of the ages of their producers. They provide a glimpse into an excited young mind as it turns counting, a boring subject for any age, into artistic gliding, with a histogram as its landing zone. This is vivid from Kiley's drawing, which depicts the process of tallying, sorting, and binning, before turning them into a histogram with flying colors, so to speak. It is also a pleasant surprise that the raw data were presented in three forms, in sticks, by numbers, and as crosses, where the last one seems to document a confirmatory exercise. Whereas we can never be sure what went through the young

[9] `https://www.thatsmags.com/china/post/29655/watch-chinese-students-use-hand-swinging-technique-at-math-competition`




(a) (b)

Figure 5: Examples of a kidstogram, which also demonstrates the participatory nature in producing them, especially the one in (a). Source: `https://orsch.net/` ("orsch" stands for *One Room Schoolhouse*)

mind, this rich kidstogram regales us with a colorful story of Kiley's engagements with the data collection, data processing, and data visualization. The participatory nature of collecting data from classmates has an added benefit of encouraging effective communication as the developing brains shape themselves through social and peer interactions.

Eli's kidstogram is simultaneously a minimalist's rendering and Picasso-esque rearranging of the same process, though I doubt Eli had any training in either style. The almost monolithic color and somewhat frosty hue perhaps were not accidental, considering the inquiry here was on the number of years spent on frozen slopes. Together, they seem to paint a rather static picture of a cold histogram, with some child-play decorations. However, once Eli's creation is viewed from a perspective orthogonal to our usual angle—literally and figuratively—a kidstogram-in-kidstogram appears, with a cleverly *figured* ramp (i.e., the disfigured number 4 over a backward and slanted number 3) leading to a piste, nicely tracking our free-falling imagination. (If a reader has trouble in picturing a sideways histogram, the reader is reminded that Picasso was not known to respect geometric or numerical proportionality.) "Awesome!" is indeed the most appropriate A-grade here, because the Picasso-esque sideways depiction reminds us of the most critical question about data science—what do the data measure?

Whereas it is likely that I have overfitted the pictorial data to my belief, it is also more likely that neither Kiley nor Eli would engage in their projects nearly as much if they were asked to simply calculate the average of their data or alike. The participatory (and pictorial) nature of forming a kidstogram may help to convert the fear for distributional thinking into a desire for social activities that with tangible cognitive and intellectual benefits, even though the children may not perceive their experiences in those adult terms. Indeed, as long as we start such activities early, the fear would not be formed in the first place or at least not to the degree to be singled out by many developing brains as particularly intimating. For example, interacting with classmates to poll their experiences and being polled by them could be challenging initially for children who suffer from various degrees of social anxiety disorders. However, the reciprocal nature of such activities can be both self-motivating and (mutually) therapeutically, especially with sensible pairing and under proper guidance of teachers with good knowledge of child psychology. Skillful educators can also engage students by guiding them to imagine the impact of different question wording, or even experiment with them to test their ideas empirically.

For example, instead of using "How many times a day do you sharpen a pencil?", ask Kiley what would happen if she changes the question to "How many times do you sharpen a pencil?" If Kiley has difficulty appreciating the difference, the teacher can engage her by asking how would she answer the new question herself. Would she provide the same answer as to the original one? Why or why not? Is the new question harder to answer than the original one? Why or why not? Is it a better question for data collection purposes than the original one? Why or why not? Such questions serve multiple pedagogical purposes. They will help young minds to appreciate the power of words and the importance of communication. They will facilitate the development of the understanding that the concept of data is fundamentally different from the concept of number, a vital distinction that our education system fails to stress. They will also demonstrate that data collection starts with its purpose, and that variations in data go beyond the differences in numbers. A young mind may not fully digest all the implications or even appreciate all the questions, just as a child may not appreciate all the rules of grammar when learning a language. But just as exposing children to a culture and expressive environment will greatly expedite their language learning, surrounding them with a data environment and a culture of thinking beyond numbers will go a long way toward engaging them in distributional thinking and ultimately developing an acumen in dealing with uncertainty, whether for risk assessments or for prediction and inference.

Indeed, there are many ways to engage generations of developing brains, such as via games or stories from children's books, employed as a part of ASA's initiatives and strategies to engage K-6 students (Martinez and LaLonde, 2020) and adopted by *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II)* issued by ASA and the National Council of Teachers of Mathematics (Franklin and Bargagliotti, 2020; Perez et al., 2021). Producing kidstograms is merely one more such activity, but one that is designed explicitly to seed the concept and habit of distributional thinking as a part of our future generations' native scientific language. In doing so, it is important to relate distributional thinking with personal welfare and decision making, as a way to continuously reinforce the learning incentivized by personal experiences. In that regard, Stephen Jay Gould's touching essay[10], "The Median Isn't the Message" (Gould, 1985, 2013), is a life-saving (and life-changing) reading on distributional thinking that we should recommend to every student.

# 6 Let's Help the Data Science Ecosystem Evolve Healthily

Each of the four proposals in this article may unnerve some of us. If so, my mission is partially accomplished, for being radical means to touch nerves or at least to push ourselves out of our comfort zones. But improving the reliability of scientific studies will always be a work in progress, and hence the specific proposals here are merely a few stepping stones toward directions for general efforts to be made by various stakeholders and communities of the scientific enterprise. The plural form of "communities" is meant to remind us that the statistical community, however large or strong, is only one of many that we statisticians should care about, if we want to sustain our co-leading roles in data science, which has ascended rapidly to become central to human n inquiries. The trio of articles in the inaugural issue of HDSR, namely the conception of data by a philosopher (Leonelli, 2019), the data life cycle by a computer scientist (Wing, 2019), and the after-life of data by an information scientist (Borgman, 2019), vividly demonstrate the vastness of both the data science topics and the data science citizenry.

[10] `https://journalofethics.ama-assn.org/article/median-isnt-message/2013-01`

Coming with this enormity is the greatly increased varieties and complexity of the problems we need to deal with. For statisticians, issues such data privacy and a host of "algorithm politics" (e.g., algorithm accountability, fairness, interpretability, transparency, trustworthiness, etc.) provides exciting newer or bigger challenges. See for example a host of articles published in *HDSR*, such as those on differential privacy (Oberski and Kreuter, 2020; Hawes, 2020), algorithm trustworthiness (Spiegelhalter, 2020), algorithm fairness (Romano et al., 2020), algorithm transparency (Rudin and Radin, 2019; Rudin et al., 2020), etc. These challenges should compel us to work harder and more creatively to maintain and enhance the quality of our work, especially when there are increasingly more reasons and incentives for rushing our studies (e.g., lack of time or other resources).

Whereas my four "radical" proposals were originally made at the ASA symposium focusing on significance tests, they have general implications for the much broader data science enterprise. Specifically, doubling variance nudges researchers to follow the time honored approach for earning trust: deliver on promises. Dirtifying Bayes is about enhancing and enlarging the community of citizen (data) scientists, a scalable force for detecting and deterring unreliable studies. Devouring pufferfish/selfish reminds policy makers and alike the effective role of incentives in encouraging reliable studies. Last and most importantly, drawing a kidstogram is about addressing the issue of unreliable studies in the most fundamental and sustainable way, i.e, via education. Or in the words of McNutt (2020), "Self-correction by design", that is, maintaining and enhancing the ability of science to self-correct by integrating pertaining training in our curriculum designs.

With data science evolving as an artificial ecosystem (Meng, 2019), harmful mutations are inevitable. Any effort to enhance scientific reliability can help the ecosystem to evolve healthily, or at least help to prevent it from serious suffering. Just as physical exercise is an effective but demanding way to keep ourselves healthy, the directions for improving scientific reliability discussed in this article and in many others (e.g., Fineberg et al., 2020; Benjamini, 2020; Bush et al., 2020; Goeva et al., 2020; Howell, 2020; Junk and Lyons, 2020; Lin, 2020; Parashar, 2020; Plant and Hanisch, 2020; Vilhuber, 2020; Willis and Stodden, 2020) all involve hard work and unremitting effort. My call to readers is therefore to help in any way you can, from developing more principled corner cutters to designing more enlightening pedagogical materials for young minds to internalizing distributional thinking.

And of course do not forget to take the "selfish oil" with every study, to help keep our collective professional body in top shape—and yours too. Thank you, my friend (or foe).

# A Proof of Theorem 1

**(I)** Given $B$ as defined in (3.1) and $B_o$ of (3.5), we have

$$\delta_o \equiv \frac{B_o}{B} - 1 = \frac{p + f_+\eta}{p + f_o} - 1 = \frac{f_+\eta - f_o}{p + f_o} = \frac{f_-}{1 - f_-} \cdot \frac{f_o - p}{p + f_o}, \tag{A.1}$$

where the last equality is due to the identity

$$(1 - f_-)(f_+\eta - f_o) = f_-(f_o - p), \tag{A.2}$$

which can be verified directly as both sides are equal to $f_-(1 - f_-)(f_+\eta - p)$. Expression (A.1) clearly implies that $B \le B_o$ if and only if $p \le f_o$. Furthermore, because $|a - b| \le |a| + |b|$, the rightmost expression of $\delta_o$ in (A.1) immediately implies that $|\delta_o| \le f_-/(1 - f_-) = O_-$ for all $p \in [0, 1]$.

We note that $p \le f_o$ if and only if $f_+\eta \ge p$, which holds if and only if $B \le 1/2$. This also implies that $B = 1/2$ if only if $f_o = p$, that is, the total error rate is the same as prevalence rate.

**(II)** Similarly, by the definition of $B_+$ of (3.7), and recalling that $\eta = (1-p)/(1-f_-)$, we have

$$\delta_+ = \frac{B_+ - B}{B} = \frac{p + f_+\eta}{p + f_+} - 1 = \frac{f_+}{p + f_+}(\eta - 1) = \frac{1 - B_+}{1 - f_-}(f_- - p). \tag{A.3}$$

This implies immediately that $B \leq B_+$ if and only if $p \leq f_-$. For the last expression in (A.3), when $p \leq f_-$, it is smaller than $O_-$, which is reached when $p = 0$. When $p \geq f_-$, its magnitude is an increasing function of $p$, and hence it reaches its maximum $f_+/(1+f_+)$ when $p = 1$. Hence, the bound in (3.8) holds for any $p \in [0, 1]$.

**(III)** Noting that $B_-$ amounts to substituting $f_+$ by $f_-$, we have

$$\delta_- = \frac{B_-}{B} - 1 = \frac{(1-p)(f_+ - f_-)}{(1-f_-)p + (1-p)f_+} = \frac{(f_+ - f_-)}{(1-f_-)O_p + f_+}. \tag{A.4}$$

All results then follow.

# References


Anup Amatya, Dulal Bhaumik, and Robert D Gibbons. Sample size determination for clustered count data. *Statistics in Medicine*, 32(24):4162–4179, 2013.

Gary S Becker. *The economic approach to human behavior*, volume 803. University of Chicago Press, 1976.

Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical significance. *Nature Human Behaviour*, 2:6–10, 2018.

Yoav Benjamini. Selective inference: The silent killer of replicability. *Harvard Data Science Review*, 2(4), 12 2020. doi: 10.1162/99608f92.fc62b261.

Michael R Berthold. What does it take to be a successful data scientist? *Harvard Data Science Review*, 1 (2), 2019. doi: 10.1162/99608f92.e0eaabfc.

Dulal K Bhaumik, Anindya Roy, Subhash Aryal, Kwan Hur, Naihua Duan, Sharon-Lise T Normand, C Hendricks Brown, and Robert D Gibbons. Sample size determination for studies with repeated continuous outcomes. *Psychiatric Annals*, 38(12), 2008.

Dulal K Bhaumik, Anindya Roy, Nicole A Lazar, Kush Kapur, Subhash Aryal, John A Sweeney, Dave Patterson, and Robert D Gibbons. Hypothesis testing, power and sample size determination for between group comparisons in fMRI experiments. *Statistical Methodology*, 6(2):133–146, 2009.

Joseph K Blitzstein and Jessica Hwang. *Introduction to probability*. Chapman and Hall/CRC, 2019.

Christine L. Borgman. The lives and after lives of data. *Harvard Data Science Review*, 1(1), 7 2019. doi: 10.1162/99608f92.9a36bdb6.

Valerie C. Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. Are We There Yet? Big Data Significantly Overestimates COVID-19 Vaccination in the US. 2021. arxiv preprint arxiv.org/abs/2106.05818.

Jackie Burt. *Orsch...Cutting the Edge in Education: Lessons Learned from an Innovative Lab School*. Stone Press, 2014.

Rosemary Bush, Andrea Dutton, Michael Evans, Rich Loft, and Gavin A. Schmidt. Perspectives on data reproducibility and replicability in paleoclimate and climate science. *Harvard Data Science Review*, 2(4), 12 2020. doi: 10.1162/99608f92.00cd8f85.

Daniel Came. Disinterestedness and objectivity. *European Journal of Philosophy*, 17(1):91, 2009.

Colin F Camerer and George Loewenstein. Behavioral economics: Past, present, future. In *Advances in Behavioral Economics* (chapter one). 2004. doi: 10.1515/9781400829118.

Shomesh Chaudhuri, Andrew W Lo, Danying Xiao, and Qingyang Xu. Bayesian adaptive clinical trials for anti-infective therapeutics during epidemic outbreaks. *Harvard Data Science Review*, Special Issue 1, 2020. doi: 10.1162/99608f92.7656c213.

John Copas and Shinto Eguchi. Local model uncertainty and incomplete-data bias (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):459–513, 2005.

David Roxbee Cox. *Principles of statistical inference*. Cambridge University Press, 2006.

David Roxbee Cox and Christl A Donnelly. *Principles of applied statistics*. Cambridge University Press, 2011.

David Roxbee Cox and Eleanor J Snell. *Applied statistics: principles and examples*. Routledge, 2018.

Robert E Fay. When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 227–232, 1992.

Usama Fayyad and Hamit Hamutcu. Toward foundations for data science and analytics: A knowledge framework for professional standards. *Harvard Data Science Review*, 2(2), 2020. doi: 10.1162/99608f92.1a99e67a.

Harvey Fineberg, Victoria Stodden, and Xiao-Li Meng. Highlights of the US National Academies Report on "Reproducibility and Replicability in Science". *Harvard Data Science Review*, 2(4), 10 2020. doi: 10.1162/99608f92.cb310198.

Christine Franklin and Anna Bargagliotti. Introducing GAISE II: A guideline for precollege statistics and data science education. *Harvard Data Science Review*, 2(4), 10 2020. doi: 10.1162/99608f92.246107bb.

Aleksandrina Goeva, Sara Stoudt, and Ana Trisovic. Toward reproducible and extensible research: From values to action. *Harvard Data Science Review*, 2(4), 12 2020. doi: 10.1162/99608f92.1cc3d72a.

Ruobin Gong and Xiao-Li Meng. Judicious judgment meets unsettling updating: Dilation, sure loss and Simpson's paradox (with discussions). *Statistical Science*, 36(2):169–214, 2021.

Stephen Jay Gould. The median isn't the message. *Discover*, 6(6):40–42, 1985.

Stephen Jay Gould. The median isn't the message. *AMA Journal of Ethics*, 15(1):77–81, 2013.

Laura Haas, Alfred Hero, and Robert A Lue. Highlights of the National Academies Report on" Undergraduate Data Science: Opportunities and Options". *Harvard Data Science Review*, 1(1), 2019. doi: 10.1162/99608f92.38f16b68.

Ian Hacking. *Logic of statistical inference*. Cambridge University Press, 2016.

Michael B Hawes. Implementing differential privacy: Seven lessons from the 2020 United States census. *Harvard Data Science Review*, 2(2), 4 2020. doi: 10.1162/99608f92.353c6f99.

Daniel F Heitjan and Donald B Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4): 2244–2253, 1991.

Emily L Howell. Science communication in the context of reproducibility and replicability: How nonscientists navigate scientific uncertainty. *Harvard Data Science Review*, 2(4), 12 2020. doi: 10.1162/99608f92.f2823096.

John PA Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

Michael Isakov and Shiro Kuriwaki. Towards principled unskewing: Viewing 2020 election polls through a corrective lens from 2016. *Harvard Data Science Review*, 2(4), 11 2020. doi: 10.1162/99608f92.86a46f38.

Thomas R Junk and Louis Lyons. Reproducibility and replication of experimental particle physics results. *Harvard Data Science Review*, 2(4), 2020. doi: 10.1162/99608f92.250f995b.

Eugene Katsevich and Aaditya Ramdas. Towards "simultaneous selective inference": Post-hoc bounds on the false discovery proportion. *arXiv preprint arXiv:1803.06790*, 2018.

Lukas Koch. Robust test statistics for data sets with missing correlation information. *Physical Review D*, 103(11):113008, 2021.

Eric Kolaczyk, Haviland Wright, and Masanao Yajima. Statistics Practicum: Placing 'practice' at the center of data science education (with discussions). *Harvard Data Science Review*, 3(1), 2021. doi: 10.1162/99608f92.2d65fc70.

Phillip S Kott. A paradox of multiple imputation. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 380–383, 1995.

Daniel Lakens, Federico G Adolfi, Casper J Albers, Farid Anvari, Matthew AJ Apps, Shlomo E Argamon, Thom Baguley, Raymond B Becker, Stephen D Benning, Daniel E Bradford, et al. Justify your alpha. *Nature Human Behaviour*, 2(3):168–171, 2018.

Sabina Leonelli. Data governance is key to interpretation: Reconceptualizing data in data science. *Harvard Data Science Review*, 1(1), 7 2019. doi: 10.1162/99608f92.17405bb6.

Xihong Lin. Learning lessons on reproducibility and replicability in large scale genome-wide association studies. *Harvard Data Science Review*, 2(4), 12 2020. doi: 10.1162/99608f92.33703976.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Wendy Martinez and Donna LaLonde. Data science for everyone starts in kindergarten: Strategies and initiatives from the american statistical association. *Harvard Data Science Review*, 9 2020. doi: 10.1162/99608f92.7a9f2f4d.

Marcia McNutt. Self-correction by design. *Harvard Data Science Review*, 2(4), 12 2020. doi: 10.1162/99608f92.32432837.

Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 9(4):538–558, 1994a. URL `http://www.jstor.org/stable/10.2307/2246252`.

Xiao-Li Meng. Posterior predictive $p$-values. *The Annals of Statistics*, 22(3):1142–1160, 1994b.

Xiao-Li Meng. AP Statistics: Passion, Paradox, and Pressure (Part I). *Amstat News*, (December):7–10, 2009a.

Xiao-Li Meng. Desired and feared — what do we do now and over the next 50 years? *The American Statistician*, 63(3):202–210, 2009b.

Xiao-Li Meng. AP Statistics: Passion, Paradox, and Pressure (Part II). *Amstat News*, (January):5–9, 2010.

Xiao-Li Meng. You want me to analyze data I don't have? Are you insane? *Shanghai archives of psychiatry*, 24(5):297, 2012.

Xiao-Li Meng. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12(2):685–726, 2018.

Xiao-Li Meng. Data science: An artificial ecosystem. *Harvard Data Science Review*, 1(1), 7 2019. doi: 10.1162/99608f92.ba20f892.

Xiao-Li Meng. Reproducibility, replicability, and reliability. *Harvard Data Science Review*, 2(4), 10 2020a. doi: 10.1162/99608f92.dbfce7f9.

Xiao-Li Meng. Information and uncertainty: Two sides of the same coin. *Harvard Data Science Review*, 2 (2), 4 2020b. doi: 10.1162/99608f92.c108a25b.

Sendhil Mullainathan and Richard H Thaler. Behavioral economics. Technical report, National Bureau of Economic Research, 2000.

Daniel L. Oberski and Frauke Kreuter. Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review*, 2(1), 1 2020. doi: 10.1162/99608f92.63a22079.

Manish Parashar. Leveraging the national academies' reproducibility and replication in science report to advance reproducibility in publishing. *Harvard Data Science Review*, 2(4), 12 2020. doi: 10.1162/99608f92.b69d3134.

Particle Data Group. Review of particle physics. *Progress of Theoretical and Experimental Physics*, 2020 (8):1–2093, 2020.

Leticia R Perez, Denise A Spangler, and Christine Franklin. Engaging young learners with data: Highlights from GAISE II, level A. *Harvard Data Science Review*, 4 2021. doi: 10.1162/99608f92.be3c2ec8.

Anne L Plant and Robert J Hanisch. Reproducibility in science: A metrology perspective. *Harvard Data Science Review*, 2(4), 12 2020. doi: 10.1162/99608f92.eb6ddee4.

Nancy Reid and David R Cox. On some principles of statistical inference. *International Statistical Review*, 83(2):293–308, 2015.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2(2), 4 2020. doi: 10.1162/99608f92.03f00592.

Anindya Roy, Dulal K Bhaumik, Subhash Aryal, and Robert D Gibbons. Sample size determination for hierarchical longitudinal designs with differential attrition rates. *Biometrics*, 63(3):699–707, 2007.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987.

Cynthia Rudin and Joanna Radin. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2), 11 2019. doi: 10.1162/99608f92.5a8a3a3d.

Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1), 3 2020. doi: 10.1162/99608f92.6ed64b30.

Alain Samson. The behavioral economics guide 2016 (with an introduction by Gerd Gigerenzer), 2016. http://eprints.lse.ac.uk/66934/7/Samson_Behavioural%20economics%20guide_%202016_author.pdf.

Nathan Sanders. Can the coronavirus prompt a global outbreak of "distributional thinking" in organizations? *Harvard Data Science Review*, 2(2), 2020. doi: 10.1162/99608f92.a577296b.

Martijn J Schuemie, M Soledad Cepeda, Marc A Suchard, Jianxiao Yang, Yuxi Tian, Alejandro Schuler, Patrick B Ryan, David Madigan, and George Hripcsak. How confident are we about observational findings in healthcare: A benchmark study. *Harvard Data Science Review*, 2(1), 2020. doi: 10.1162/99608f92.147cc28e.

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

David Spiegelhalter. Should we trust algorithms? *Harvard Data Science Review*, 2(1), 1 2020. doi: 10.1162/99608f92.cb91a35a.

Victoria Stodden. Theme editor's introduction to reproducibility and replicability in science. *Harvard Data Science Review*, 2(4), 12 2020. doi: 10.1162/99608f92.c46a02d4.

John W Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962.

Gerald Van Belle, Lloyd D Fisher, Patrick J Heagerty, and Thomas Lumley. *Biostatistics: A methodology for the health sciences*, volume 519. John Wiley & Sons, 2004.

Lars Vilhuber. Reproducibility and replicability in economics. *Harvard Data Science Review*, 2(4), 12 2020. doi: 10.1162/99608f92.4f6b9e67.

Lance Waller and Taal Levi. Building intuition regarding the statistical behavior of mass medical testing programs. *Harvard Data Science Review*, Special Issue 1, 2021.

Ronald L Wasserstein and Nicole A Lazar. The ASA's statement on $p$-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

Nick Wilkinson and Matthias Klaes. *An introduction to behavioral economics*. Macmillan International Higher Education, 2017.

Craig Willis and Victoria Stodden. Trust but verify: How to leverage policies, workflows, and infrastructure to ensure computational reproducibility in publication. *Harvard Data Science Review*, 2(4), 12 2020. doi: 10.1162/99608f92.25982dcf.

Jeannette M Wing. The data life cycle. *Harvard Data Science Review*, 1(1), 7 2019. doi: 10.1162/99608f92.e26845b4.

Min-ge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1):3–39, 2013.

Xianchao Xie and Xiao-Li Meng. Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? (with discussion). *Statistica Sinica*, 27:1485–1594, 2017.