Statistical Science
2021, Vol. 36, No. 2, 210–214
https://doi.org/10.1214/21-STS765REJ
Main article: https://doi.org/10.1214/19-STS765

Rejoinder: Let's Be Imprecise in Order to Be Precise (About What We Don't Know)

Ruobin Gong and Xiao-Li Meng

Preparing a rejoinder is a typically rewarding, sometimes depressing and occasionally frustrating experience. The rewarding part is self-evident, and the depression sets in when a discussant has much deeper and crisper insights about the authors' thesis than the authors themselves. Frustrations arise when the authors thought they made some points crystal clear, but the reflections from the discussants show a very different picture. We are deeply grateful to the editors of *Statistical Science* and the discussants for providing us an opportunity to maximize the first, sample the second and minimize the third.

1. LET'S AUGMENT OUR SHOES TO FIT OUR GROWING FEET

Professor Glenn Shafer's historically infused and theoretically fermented insights provided us with an intense savoring and much lingering. His succinct summary of the three branches of the art of conjecture of d'Alembert laid out the contours and interplay among (precise) probability, statistics and imprecise probability. The first branch enters the game of conjecture by manipulating theoretically precisely specified quantities and models, a game of precise probability, deducing properties and consequences of a theoretical construct.

The second branch plays the same game empirically, by focusing on assessing chances and risks from data. This captures the essences of the current statistical practices, when empirical assessments are guided by the rules of precise probability. Principled statistical practices fully recognize the multiple uncertainties in empirical assessments, and hence have built-in risk assessments for estimating the part of uncertainties that can be reasonably gauged empirically. For parts that cannot be empirically assessed internally, sensitivity studies have been the primary tool, precisely because by posting specific alternative scenarios, we can traverse within the first two branches, and hence remain in our comfort zone.

Shafer's summary made clear that the third branch clamors more attention than we currently bestow. This branch covers the vast majorities of inquires where precise probabilistic descriptions, whether theoretical or empirical, are inherently incomplete or impossible. In our own applied work, when we ask a subject expert to provide a prior, the most precise answer would be of the kind "I'm quite sure that α is between 1 and 2." Any further inquiry about how α is distributed on $[1, 2]$ would be met with either a puzzled face or an answer few of us like: "I have no idea."

Such "no-idea" answers have motivated many to work harder throughout history. Nevertheless, currently we are still forced to make up assumptions, such as α is distributed uniformly on $[1, 2]$, for the sole purpose of applying available theories or methods. Or as Shafer put it, despite efforts to move bits of the third branch into the first two, "the third still seems very large."

Instead of cutting feet to fit shoes, the framework of *imprecise probability* (IP) suggests a less painful paradigm: expanding the shoes to fit the feet. This metaphor has another leg to stand on because the imprecise shoes are no less functional than the precise ones. As Augustin and Schollmeyer emphasized, IP should have been better named as "set-valued probability." But sound statistical inference is already set-valued, as classical paradigms have delivered via confidence intervals and Bayesian credible sets, in order to reflect inferential uncertainty. In that sense, the set-valued output of IP models is no less familiar a mathematical form than that from precise probability models, albeit carrying a different interpretation of "uncertainty". It is therefore natural for us to ask: why can't we go from set-valued input to set-valued output directly, instead of squeezing through the narrow tunnel of numerically valued probability?

2. TWO CONCERNS THAT MOTIVATED OUR WORK

To answer this question, we would like to elaborate our view on the role of imprecise probabilities and their accompanying updating rules. We surmise nearly all statisticians take for granted that probability is the language of uncertainty. And by probability, we specifically mean countably additive probability that obeys the Kolmogorov axioms. Bayesians, Frequentists, as well as those who entertain fiducial, structural and functional inference, all operate within a framework that guides the expression of uncertainty relating observable information to unknown

Ruobin Gong is Assistant Professor of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, New Jersey 08854, USA (e-mail: rg915@stat.rutgers.edu). Xiao-Li Meng is Whipple V. N. Jones Professor of Statistics, Harvard University, 1 Oxford Street, Cambridge, Massachusetts 02138, USA (e-mail: meng@stat.harvard.edu).

quantities of interest, and in this sense, *update* their knowledge in light of what has been learned.

Probability is the magical lasso with which statisticians tame unruly variabilities. We use probability to gloss over irregularities that we prefer not delve deeper to model. We invoke, often rightly so, some notions of independence or exchangeability, motivated by our *ignorance* on any relational information. We use ignorance to our advantage: the theory of randomization is founded on variabilities, artificially induced in a way that make us conclude that we are better off not to delve deeper. The essence of any statistical analysis relies on the judicious reduction of the unknown and the unknowable into known or knowable probabilities. One of us discussed (Meng, 2014, Liu and Meng, 2016, Li and Meng, 2021) the multiresolution nature of inference, which gives meaning to the word "judicious" in this context via the choice of the resolution level. Reduction is great—until when it has gone too far to the point of absurdity, that is, the resolution becomes too low. This is where imprecise probabilities comes in.

In general, we are concerned with the change of the "model" $P(A)$ to $P_B(A)$, after B has been learned. We borrow here Hausdorff's notation for relative probability, introduced by Shafer. The Bayes rule, namely the assertion that upon observing event B, the agent shall replace her prior belief $P(A)$ about A with the conditional probability $P(A \mid B)$, has been justified as the rule to instill such a change, even in the context that such learning is to happen dynamically over time. Carnap (1962) called $\mathrm{Cr}_{X,T}(H)$ the degree of credence of agent X toward assertion H at time T, and $\mathrm{Cr}_{X,T}(H \mid E)$ the conditional credence if X ascertained that proposition E holds. He maintained that if E is the observational data received by X between times t_1 and t_2, the rational and coherent agent X must transform their credence at time t_1 to time t_2 as $\mathrm{Cr}_{X,t_2}(H) = \mathrm{Cr}_{X,t_1}(H \mid E)$. Teller (1973)'s dynamic Dutch book argument attempted to compel the same conclusion. He argued that if an agent engages in a mixture of regular and called-off bets at prices that differ from their assessed marginal and conditional probabilities of the uncertain outcome, they would be made a sure loser by the exploitative (and know-it-all, we must add) bookie.

These arguments, although compelling, were applicable only within a limited and highly idealized scope. To dynamically update one's credence via Bayesian conditionalization requires that the agent knows the "full road map" ahead of time, which typically is not the case in practice. This is why, first and foremost, we celebrate the potential of IP tools in resolving this matter. We are therefore happy to see the support and enhancement by the discussants from multiple vantage points. Augustin and Schollmeyer provided us with a succinct overview of the use of credal sets in statistical inference, which is appealing because it is rooted in distributional families, a concept familiar to statisticians. Liu and Martin, here and more generally in their work on inferential models (IM; Martin and Liu, 2015), argued that to achieve the validity as they defined, we must resort to IP models. Wheeler opened our eyes further by introducing us to a world with constructs that are even more primitive than credal sets.

Nevertheless, invoking IP quantification resolves only one aspect of the overly aggressive reduction to uncertainty reasoning that requires precise probabilities. In many situations, the statistician regardless of persuasion must contemplate how to update knowledge, not in terms of uncertainty but rather *in the presence* of uncertainty. By *in the presence*, we mean that the analyst is not certain of what model structure they are to construct in the first place, or that they do not have an idea how isolated pieces of information, such as individual observations, interact with each other. In the terminology of Liu and Martin, it is the "association equation" that is ill-defined. This is why the focus of our article is not on the IP description itself, but rather conditioning (and by extension, combining) rules for IP. For example, what is our equivalent go-to assumption, like exchangeability in the precise world?

The examples presented in our article were chosen for their simple nature. Thus, when disagreements arise among the rules in question, not only is their effect unmistakably stark, but also the reader may appeal to her/his intuitive judgment as to which answers are more sensible than others. The IM treatment offered by Liu and Martin works well in these situations, because it is known to the modeler *ex ante* what kind of inferential conclusion is more desirable. Their model building process, including the specification of the association and the choice of the predictive random set, reflects these convictions of the modeler and produces—without surprise—results that are both desirable and intuitive. Pedagogical examples can go only so far, however. The leap from simple examples to reality deprives the modeler the luxury of intuition, and the decision on how to update becomes nontrivial when it is no longer preceded by the answer. Just like other modes of IP-based frameworks of inference, the IM framework is faced with a nontrivial choice of rules when it comes to combining marginal models (Martin and Syring, 2019). A fundamental question is whether this choice shall be guided by so-called desirable properties that pertain to the resulting answer, knowing that our notion of what properties are desirable is riddled with inaccurate assumptions, which IP models aim to address in the first place.

3. PARTIAL ORDERING: IS IT A FEATURE OR A BUG?

Mathematically, the multiplicity of rules is a result of using sets to capture the low-resolution nature of our data or information (Gong and Meng, 2021). Sets obey only partial ordering: a set A can be neither larger nor smaller than another set B. Just as ambiguities in life typically lead to multiple scenarios, partial ordering permits multiple ways to revise our probabilistic assessments to take into account additional considerations. We therefore argue partial ordering necessitated by treating sets as the fundamental building blocks for probability specifications is rather a feature, not a bug, for dealing with imprecise data, information, or other forms of inconclusive evidence.

Our article focused mostly on studying and comparing individual rules, instead of seeking a deeper unification. We are therefore grateful to Augustin and Schollmeyer for giving us a healthy dose of depression, as their elegant envelope representation is the unification we missed when we attempted to discuss the "optimism" of Dempster's rule, the "pessimism" of the Geometric rule, and the "conservatism" of the generalized Bayes rule, in the three-prisoner example. Their envelope representation makes it crystal clear that (1) ideological differences are inherently embedded into the rules, hence are omnipresent; and (2) behavioral differences among the rules are driven by their underlying ideology.

The envelope representation formulates all three rules as obtaining extreme probabilities subject to different constraints in a family of distributions. The generalized Bayes rule is the most conservative because it assumes no (further) constraint, resulting the widest possible probability interval $[\underline{P}, \overline{P}]$. To better understand the optimism of Dempster's rule and the pessimism of the Geometric rule, it helps to consider the case of belief function, where we can map a set of probabilities to an ordinary probability of sets (e.g., Gong and Meng, 2021).

Under the precise probability formulation, when we reassess a probability by moving from its original state space Ω to a subset $S \subset \Omega$, we will permit (and permit only) any $\omega \in S$. In contrast, for belief function, "moving from Ω to B" can have multiple interpretations due to the ambiguity from partial ordering. We can take the most generous route by permitting any (nonzero mass) set $A \subseteq \Omega$ that is not ruled out by B, that is, any $A \cap B \neq \varnothing$. This is the route that Dempster's rule takes, an optimistic choice since $A \cap B \neq \varnothing$ permits (far) more states ω than $A = B$ would. In contrast, the Geometric rule permits only any $A \subseteq B$. That is, a state ω (and its parental set A) is permitted only if it is in B, hence the most pessimistic—or putting it more positively—the most cautious route. One can also "refuse" to make a choice by seeking extremes over all the rules, but that merely means adopting the generalized Bayes rule, which in our view brings another kind of trouble, as Section 5 will discuss.

4. ARE WE TOO PESSIMISTIC?

Wheeler and Liu-Martin cast their discussions from very different perspectives. Wheeler's supplied rich background from the IP literature accompanied by a logician's rigor and insight. Liu and Martin took an operational perspective with an utilitarian flavor. They, however, reached essentially the same conclusion, that our article projects a sense of pessimism by (overly) emphasizing "intrinsic contradictions" within the IP paradigm. Wheeler pointed out that we missed the entire contemporary theory of *lower previsions*, which includes lower probabilities as a special case, and where "coherence preservation under inference is inviolable." Liu and Martin criticized us for not imposing criteria of reliability, which could cure or at least reduce our unsettling feeling.

Wheeler was entirely correct that we missed the theory of lower previsions. We are frustrated by our ignorance, and the long learning curve, reflecting Shafer's observation that "the theory of imprecise probability has flourished for several decades, but largely outside of statistics journals." We therefore particularly appreciate the editors of *Statistical Science* for seeing the value of this topic and for organizing this discussion, which also provides us with a great learning opportunity.

Liu and Martin were also correct that we did not explicitly impose any reliability criteria. In the sentence they quoted, we made it explicit that the rules are "prespecified," and the choice of criteria is a part of the prespecification. To us, precise probability is the grammar for statistical inference under the highest-resolution specifications, that is, when we can—or pretend we can—postulate probability specifications on all individual elements in however complicated or high-dimensional joint state spaces. Adopting the Bayes rule as suggested by the Bayes theorem, a consequence of precise probability, can be viewed as a reliable, criteria-driven exercise (e.g., by imposing the coherence requirement). But as Shafer correctly pointed out, the distinction between Bayes rule and Bayes theorem has been essentially ignored in the statistical literature (and we certainly accept Shafer's criticism for our own "confusing" mix of the two). We surmise this was largely due to acceptance of precise probability as a *reliable* grammar for statistical inference, and hence any rules set or implied by it would be accepted without the need for further criteria to justify them.

An initial motivation for our work was our desire to find out the natural *generalization* of the Bayes rule, as given by Shafer's (1), in the world of imprecise probabilities. The singular form of "generalization" was intended, as for decades, one of us believed (or hoped) that Dempster's rule was *the* natural generalization of the Bayes rule, implied by some "Dempster theorem," a consequence of the belief function apparatus. The dream was broken when the other of us actually studied the issue (not just dreamt about it), and what we presented was a part of that broken dream.

We therefore hoped that we were wrong, and that our "pessimism" was a result of our ignorance, that is, we had not looked hard enough. Consequently, we were excited initially by Wheeler's emphasis that what we explored only reflected what was known before either of us was born. Whereas we definitely want to and will study the more contemporary theory of lower previsions, a more careful reading of Wheeler's discussion reignited our unsettling feeling because the lower previsions preserve only the coherence, and Wheeler's conclusion is that the generalized Bayes rule is still his preference. That is, our pessimism is not a reflection of our ideology, but fundamental to the marriage of coherence with IP, as we explain below.

5. SHOULD WE ALSO AVOID FATAL ATTRACTION?

If coherence were the only desirable criterion, the generalized Bayes rule would indeed be the choice. However, the generalized Bayes rule suffers from a flaw that is no less serious or fatal than being incoherent, that is, it cannot get itself out of the vacuous state of knowledge, regardless of the amount of data or information one accumulates. In other words, the vacuous state is a fatal attraction state of the generalized Bayes rule. We want to emphasize that the "vacuous state" is not a strawman. Much of "objective Bayes" or fiducial inference hopes to conduct distributional inference without imposing any prior knowledge, that is, to start with the vacuous state. Any updating rule that has the vacuous state as its fatal attraction clearly will be eliminated from the start.

Similarly, the fact that Liu and Martin's "validity" requirement can lead to the vacuous state as the only solution (e.g., see Section 4.3 of Liu and Martin's discussion) raises the question of the general desirability or even the validity of this "validity" requirement. Indeed, Liu–Martin's validity requirement is fundamentally a frequentist calibration construct, like unbiasedness for testing. It inherits the known defects of their classical counterparts (e.g., controlling Type I error), such as a lack of relevance, or making the wrong trade-off by assigning higher confidence to harder problems (e.g., Liu and Meng, 2016).

We raise these points not to suggest that we have better solutions, but to reaffirm that judicious judgment and choices are inevitable. Inference is not possible without making assumptions, and any assumption is a judgmental call, judicious or not. The dominant emphasis on coherence in the IP literature suggests that itself is a choice. The generalized Bayes rule is the only coherent updating rule for sets of probabilities in the sense of avoiding sure loss. However, it is incompatible with other sensible considerations, such as avoiding fatal attraction of the vacuous state, posing an intrinsic contradiction that we must accept.

6. THE DEMAND (AND SUPPLY) FOR JUDICIOUS JUDGMENT

Even in the precise probability situation, the analyst often does not know what model to specify, except that their partial and meta-knowledge makes them realize that the Bayesian recipe of conditionalization may *not* be the best course of action under the model they are forced to posit. An example of this is on modularized Bayesian inference and cut distributions (Liu, Bayarri and Berger, 2009, Lunn et al., 2009, Plummer, 2015, Jacob et al., 2017), inspired by Bayesian pharmacokinetics and pharmacodynamics (PKPD) models. The analyst has information that a certain margin of the joint model may involve poor quality data or information, and would like to sever the contribution of this margin to other parts of the model for which the analyst has scientific interest. In theory, if we know how to quantify data or information quality, we can incorporate such quantification properly in our probabilistic model. Following the Bayesian recipe, such as conditioning, would lead to sensible inference that properly weights various pieces of information by their quality index. However, other than for linear estimates and estimators (Meng, 2018), quantifying deterioration in quality due to nonsampling mechanisms is currently infeasible.

A common practical approach is then to attach zero weight to the problematic aspects of the data or model, that is, to "cut them off." This is often a better strategy than keeping them, because zero weight is likely a better approximation to the (unknown) optimal weights than blindly pretending that all parts of the data or model should be given the standard treatments, for example, equal weighting. Evidently, such "updating" procedures through cut distributions do not conform to Bayesian conditionalization. However, they tend to yield better results in practice, because they are better approximations to the optimal but inoperable Bayesian approach under the fully correctly specified model, than mechanically applying the Bayesian rule to the misspecified model.

We therefore thank Augustin and Schollmeyer for their proposal of soft revision, as a customizable updating rule that bridges the pessimist Geometric rule and the optimist Dempster's rule, to which they drew a connection to the maximum Bayes factor approach of Good (1967) and an analogy with empirical Bayes. We were reminded of Lindley's declaration that "there is no one less Bayesian than an empirical Bayesian" (Lindley, 1969, in discussion of Copas, 1969). As much as the proposed soft revision ventures outside the realm of coherent Bayesianism, it is a useful and welcome addition to the toolbox of the practical statistician, one that could help us avoid the fatal attraction.

As we encourage the use of imprecise probabilities, there is risk in harming the operationalizability of the statistical inference framework. We view this as yet another instance of the omnipresent no-free lunch principle. Indeed, there is a price to pay even for every precise generalization of the ordinary probability calculation. Good (1966) advocated for *probabilities of higher types* as a candidate measure of nonmeasurable events. He remarked, however, that probabilities of higher types are expressible only in terms of inequalities that are fuzzy in nature, and quickly lose practical importance the higher in type they go. Similarly, if we were to jettison the notion of a relatively well-defined protocol, and possibly other aspects of routine practice of model building, it would not be long before a necessary level of operationalizability is lost.

As widely endorsed as Bayesian analysis is among applied statisticians, its computational challenge once was considered insurmountable, until the MCMC revolution brought the change. Customizable apparatuses, such as WinBUGS and Stan, made Bayesian computation on large-scale datasets widely accessible. By way of contrast, computation for IP models in statistical inference is still in its early development. The SIPTA (Society for Imprecise Probability: Theories and Applications) community has seen recent advances on the use of MCMC to estimate lower expectations (Fetz, 2019, Decadt, de Cooman and De Bock, 2019). The statistical literature is starting to catch up in that regard. The recent work of Jacob et al. (2021) developed the first workable sampler for the random convex polytope characterizing the Dempster–Shafer inference for categorical data, proposed 50 years prior (Dempster, 1966, 1972).

The motto of the SIPTA community is that *there are more uncertainties than probabilities*. As statisticians, we are eager to see this reflected in the practice of statistical inference. By conducting imprecise probability inference, we can be precise about what we do not know, and hence deliver more replicable results because we avoid making up assumptions forced upon us by the precise probability framework. We therefore invite anyone who cares about scientific replicability to look into what the world of IP can offer.

REFERENCES

CARNAP, R. (1962). *Logic, Methodology and Philosophy of Science. Proceedings of the* 1960 *International Congress*. Stanford Univ. Press, Stanford, CA. MR0166069

COPAS, J. B. (1969). Compound decisions and empirical Bayes. *J. Roy. Statist. Soc. Ser. B* **31** 397–417. MR0269013

DECADT, A., DE COOMAN, G. and DE BOCK, J. (2019). Monte Carlo estimation for imprecise probabilities: Basic properties. In *International Symposium on Imprecise Probabilities: Theories and Applications. Proc. Mach. Learn. Res.* (*PMLR*) **103** 135–144. MR4020914

DEMPSTER, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Stat.* **37** 355–374. MR0187357 https://doi.org/10.1214/aoms/1177699517

DEMPSTER, A. P. (1972). A class of random convex polytopes. *Ann. Math. Stat.* **43** 260–272. MR0303650 https://doi.org/10.1214/aoms/1177692719

FETZ, T. (2019). Improving the convergence of iterative importance sampling for computing upper and lower expectations. In *International Symposium on Imprecise Probabilities: Theories and Applications. Proc. Mach. Learn. Res.* (*PMLR*) **103** 185–193. MR4020919

GONG, R. and MENG, X.-L. (2021). Probabilistic underpinning of imprecise probability for statistical learning with low-resolution information. Technical report.

GOOD, I. J. (1966). Symposium on current views of subjective probability: Subjective probability as the measure of a non-measurable set. In *Studies in Logic and the Foundations of Mathematics* **44** 319–329. Elsevier, Amsterdam.

GOOD, I. J. (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. Ser. B* **29** 399–418. MR0225428

JACOB, P. E., MURRAY, L. M., HOLMES, C. C. and ROBERT, C. P. (2017). Better together? Statistical learning in models made of modules. Preprint. Available at arXiv:1708.08719.

JACOB, P. E., GONG, R., EDLEFSEN, P. T. and DEMPSTER, A. P. (2021). A Gibbs sampler for a class of random convex polytopes. *J. Amer. Statist. Assoc.* To appear. With discussion.

LI, X. and MENG, X.-L. (2021). A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2020.1844210

LINDLEY, D. V. (1969). Discussion of compound decisions and empirical Bayes, J. B. Copas. *J. Roy. Statist. Soc. Ser. B* **31** 419–421.

LIU, F., BAYARRI, M. J. and BERGER, J. O. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Anal.* **4** 119–150. MR2486241 https://doi.org/10.1214/09-BA404

LIU, K. and MENG, X.-L. (2016). There is individualized treatment. Why not individualized inference? *Annu. Rev. Stat. Appl.* **3** 79–111.

LUNN, D., BEST, N., SPIEGELHALTER, D., GRAHAM, G. and NEUENSCHWANDER, B. (2009). Combining MCMC with 'sequential' PKPD modelling. *J. Pharmacokinet. Pharmacodyn.* **36** 19.

MARTIN, R. and LIU, C. (2015). *Inferential Models: Reasoning with Uncertainty*. CRC Press, Boca Raton, FL.

MARTIN, R. and SYRING, N. (2019). Validity-preservation properties of rules for combining inferential models. In *International Symposium on Imprecise Probabilities: Theories and Applications. Proc. Mach. Learn. Res.* (*PMLR*) **103** 286–294. MR4020929

MENG, X.-L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). In *Past, Present, and Future of Statistical Science* (X. Lin et al., eds.) 537–562.

MENG, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **12** 685–726. MR3834282 https://doi.org/10.1214/18-AOAS1161SF

PLUMMER, M. (2015). Cuts in Bayesian graphical models. *Stat. Comput.* **25** 37–43. MR3304902 https://doi.org/10.1007/s11222-014-9503-z

TELLER, P. (1973). Conditionalization and observation. *Synthese* **26** 218–258.