Statistical Science
2021, Vol. 36, No. 2, 191–195
https://doi.org/10.1214/21-STS765A
Main article: https://doi.org/10.1214/19-STS765

Comment: On the History and Limitations of Probability Updating

Glenn Shafer

Abstract. Gong and Meng show that we can gain insights into classical paradoxes about conditional probability by acknowledging that apparently precise probabilities live within a larger world of imprecise probability. They also show that the notion of updating becomes problematic in this larger world. A closer look at the historical development of the notion of updating can give us further insights into its limitations.

Key words and phrases: Bayes's rule of conditioning, Dempster's rule, conditional probability, conditionalization, imprecise probabilities, probability protocols, relative probability, updating.

1. A BROADER PERSPECTIVE ON CLASSICAL PARADOXES

Conditional probability paradoxes, stories in which $P(A|B)$ does not seem to be a reasonable probability for A after we learn B, have been with us since the late 19th century.[1] Many of these paradoxes turn on initial probabilities not telling us enough about the relation between A and the event that we learn B. Many authors have explained this, but each in their own way, often vociferously denying the cogency of others' explanations. No consensus having emerged, the paradoxes endure.

Roubin Gong and Xiao-Li Meng propose a broader perspective. Instead of trying to resolve the paradoxes within standard probability theory, in which we have joint probabilities for all events of interest, they propose that we use the theory of imprecise probabilities, in which events of interest may have only upper and lower probabilities and quantities of interest may have only upper and lower expected values. The theory of imprecise probabilities not only generalizes the standard theory but also allows us to recognize formally the incompleteness of any standard (a.k.a. "precise") probability model. We do this by adding events to the model without adding probabilities for them, thus obtaining a larger "imprecise" model. As Gong and Meng put it,

> Every precise model is a fully specified margin nested within a larger, ever-augmentable model, with extended features not allowed to enter the scene as the modeler lacks the knowledge to do so precisely.

This allows them to explain the conditional probability paradoxes this way:

> Their narratives imply the existence of a joint distribution, yet only a subset of marginal information is precisely specified.

The theory of imprecise probability has flourished for several decades, but largely outside statistics journals. Bringing it into the statistical mainstream, as Gong and Meng have done with this article in *Statistical Science*, is a welcome move. As Gong and Meng show, the theory's ideas can enrich statisticians' understanding of longstanding questions within our community. We can also hope that the critical resources of the statistical community can add new depth to the theory. Gong and Meng tell us that dilation, contraction and sure loss "hint at novel types of information contribution." Perhaps we need theories of these novel types.

2. DO ALL EVENTS HAVE NUMERICAL PROBABILITIES?

The theory of imprecise probabilities says no. Many events, perhaps most, do not have numerical probabilities. Is this a new or controversial view?

Certainly it is not new. Before the 18th century, scholars who wrote about degrees of probability seldom suggested that these degrees could ever be put in numerical form Knebel (2000). Before 1713, when Jacob Bernoulli's *Ars conjectandi* appeared, even expectations in games of chance were not usually connected with the idea of probability. Bernoulli made the connection and launched the

Glenn Shafer is University Professor, Rutgers University, Newark, New Jersey (e-mail: gshafer@rutgers.edu).

[1] See Bertrand (1889). Bertrand's paradoxes have been discussed by Shafer and Vovk (2006), Gorroochurn (2012) and many others.

project of finding numerical probabilities not only for games of chance but also for civil, criminal and business matters. But Bernoulli did not believe that we can always find probabilities for a thing and its contrary that add to one.

Jean Le Rond d'Alembert, the uncontested leader of French mathematics in his time, was an avowed skeptic about Bernoulli's ambition for numerical probabilities. In 1676, the same year the teenage Pierre-Simon Laplace arrived in Paris seeking his patronage, d'Alembert published his own views about the art of conjecture. According to d'Alembert, this art has three branches (D'Alembert, 1767, Chapter VI):

1. The first branch is games of chance. Here, we can count equally likely cases and reason about them *a priori*.[2]
2. The second consists of topics such as insurance and inoculation, where we can learn the number of cases and their ratios only from experience and only approximately.
3. The third consists of the many topics for which mathematical demonstration is rare or impossible. D'Alembert included here physics, history, medicine, the law and business.

Outside the small world of scholars who specialize in mathematical probability and its applications, these views probably found widespread assent when d'Alembert published them and may continue to do so today. Over time, scientists and statisticians may have moved bits of d'Alembert's third category into the second or even the first, but the third still seems very large.

When I began my own study of mathematical statistics in the early 1970s, I took it for granted that only some events have probabilities. Both R. A. Fisher and Andrei Kolmogorov had said so explicitly.[3] I thought nearly all statisticians, philosophers and mathematicians agreed. Today I am not so sure. For decades now, Bayesians have insisted that a person can supply a personal probability for anything. As realism has gained ascendancy in philosophy, the claim that anything uncertain has an objective probability, usually unknown, has also become common. Many physicists now imagine a universal wave function. Many mathematical probabilists now imagine the whole course of the world being described by a single element ω of a vast probability space Ω. In this context, I am tempted to see the increasing popularity of the theory of imprecise probabilities as a return to d'Alembert's common sense.

3. MODEL OR JUDGMENT?

As leaders in the "Bayesian, Fiducial and Frequentist" community, Gong and Meng want to transcend the quarrels between proponents of different interpretations of probability and different methodologies for statistical inference. This is visible in their choice of words. They avoid saying whether the probabilities they discuss, precise or imprecise, are objective facts or subjective beliefs, and they make heavy use of the word "model." The first two sentences of their article reveal, however, that the models being studied are akin to neo-Bayesians. They update themselves:

> Statistical learning is a process through which models perform updates in light of new information, according to a prespecified set of operation rules. As new observations arrive, a good statistical model revises and adapts its uncertainty quantification according to what has just been observed.

By the end of the article, however, I was wondering whether these first two sentences were a declaration of faith or a straw man. Is "judicious judgment" limited to choosing an updating rule before the fact, incorporating it into the model, and letting the model do our later thinking for us? Or is "judicious judgment" most needed after something unexpected is observed? I would welcome the second interpretation and see it as another step back to common sense.

4. FROM RELATIVE TO CONDITIONAL PROBABILITY

Two centuries before the formula

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \tag{1}$$

became a definition, Abraham de Moivre provided a betting argument for what became known as "the rule of compound probability": the probability of two events both happening is the probability of the first times the probability of the second "when the first shall have been consider'd as having happen'd" (de Moivre, 1967, p. 7). As this formulation reveals, De Moivre did not begin with a probability measure that gave joint probabilities for all events he wanted to discuss. Instead he constructed joint probabilities from simpler ingredients. The probability of a second event given the first was one of these ingredients. It was not a "conditional probability"; it was the probability of the second event in the new situation in a betting game. The rule of compound probability remained one of the basic rules of probability theory until the mid-20th

[2] D'Alembert was also skeptical about some of this *a priori* reasoning. Can you really know *a priori* the probability of getting a head tossing a coin when you are allowed to try twice? As Bernard Bru has argued, we should hesitate to dismiss d'Alembert's doubts on this point as a mere "gambler's fallacy" (Bru, 1989, 2002).

[3] Kolmogorov's most explicit statement that not every event has a probability may be in his article on probability in the 1951 edition of the *Great Soviet Encyclopedia* (Shafer and Vovk, 2006, p. 50). Fisher was equally explicit, stating in 1956, for example, that "in some cases no probability exists" (Fisher, 1956, p. 45).

century, when mathematical probabilists decided that it was more convenient to make probability measures their starting point, thus shifting (1) from being a consequence of the rule of compound probability to being a definition of $P(\cdot|\cdot)$.

Nineteenth-century mathematicians sometimes wrote about "relative probability." In his popular French textbook on probability, first published in 1816, Sylvestre-François Lacroix called the ratio $P(A)/(P(A)+P(B))$ the *probabilité relative* of A as compared to B. When rolling two dice, for example, where there are 6 chances for getting a 4 and only 3 chances for getting a 4, the probability of 7 relative to 4 is 2/3 (Lacroix, 1816, pp. 19–20). We see this same notion of relative probability in Liagre (1852), Section 16.

It seems that "conditional probability" first appeared in George Boole's *Laws of Thought* (Boole, 1854, Chapter XX, Section 21). A logician, Boole was trying to make mathematical probability part of logic, and he was accustomed to using "condition" and "conditional" in logic. Boole's used "conditional probability" only once, however, casually and perhaps even inadvertently, as he was writing mostly about "conditional events." In 1887, in his Metretike, Francis Edgeworth, citing Boole, systematically called the probability of an effect given a cause a "conditional probability" (Mirowski, 1994). We already see the German and Russian equivalents, *bedingte Wahrscheinlichkeit* and условная вероятность, in the early 20th century (Shafer and Vovk, 2006, p. 6).

In the course of commenting on Boole, Charles Sanders Peirce wrote, "Let b_{a} denote the frequency of b's among the a's" (Peirce, 1867, p. 255). Because Peirce was identifying probability with frequency, this could be considered the first notation for conditional probability. Others made other suggestions, mostly independently of each other. Hugh McColl, independently of Peirce, wrote "The symbol x_a denotes the chance that the statement a is true on the assumption that the statement a is true" (McColl, 1879/80, 1880/81). Later he used $\frac{A}{B}$ (MacColl, 1896/97). Andrei (Markov, 1900) wrote (A, B).

In 1911, John Maynard Keynes introduced what he called "the fundamental symbol of probability," A/H, for the probability of A given H. This symbol became popular at Cambridge; we see it in books by C. D. Broad (Broad, 1914, p. 318), John Maynard Keynes (Keynes, 1921, p. 177), and William E. Johnson (Johnson, 1924, p. 179). To all appearances, Keynes first used the symbol in the 1908 dissertation that grew into his book, and Johnson popularized it in conversations and lectures.[4]

In 1901, the German mathematician Felix Hausdorff introduced the symbol $P_F(E)$ for what he called the *relative Wahrsheinlichkeit von E, posito F* (relative probability of E given F). In his view, the absolute probability $P(E)$ of an event E is simply the relative probability $P_F(E)$, where F is our current knowledge. This knowledge can change, and Hausdorff mentioned three examples (Hausdorff, 1901, pp. 154–155):

- When the absolute probability $P(E)$ is a weighted average of possible objective probabilities, F represents one of the possible objective probabilities, and we learn that F is correct, then we change $P(E)$ to $P_F(E)$.
- We may learn that there were more possibilities than we had realized, as when we learn that the geometry of the world may not be Euclidean. In this case, we shift from $P_F(E)$ to $P_G(E)$, where G permits this wider set of possibilities.
- We may learn that our knowledge F was flatly wrong and, therefore, shift from $P_F(E)$ to $P_G(E)$, where G contradicts F.

Emmanuel Czuber followed Hausdorff's terminology and notation in the second edition of his influential textbook, except that he used $\mathfrak{W}_F(E)$ instead of $P_F(E)$ (Czuber, 1908, pp. 44–45). Kolmogorov used $P_A(B)$ in his pathbreaking 1933 *Grundbegriffe*, but he called such a probability *bedingte* (conditional), not *relative* as Hausdorff and Czuber had done (Kolmogoroff, 1933, p. 206).

Our current notation $P(\cdot|\cdot)$ is apparently due to Harold Jeffreys. In 1919, Dorothy Wrinch and Jeffreys had used $P(p:q)$ (Wrinch and Jeffreys, 1919). In 1931, Jeffreys replaced this with $P(p|q)$, commenting on its advantage over $P(p:q)$ and notation p/q in a way that makes clear that he was not aware of any previous use of $P(p|q)$ (Jeffreys, 1931, p. 31).

5. FROM CONDITIONAL PROBABILITY TO UPDATING

After World War II, mathematicians, statisticians and philosophers began to take it for granted that the proper setting for mathematical probability is a probability measure rather than a collection of probabilities less structured or structured in some other way. Only then did it become natural to recast the notion of conditional probability as an action with probabilities as its object: a statistician or scientist "conditionalizes" or "conditions" or "updates" the probabilities. This formulation seems to have slipped unheralded into many minds. The earliest instance of it I have found is in Estes and Suppes (1957). After emphasizing the importance for psychology of the concept of

[4] Keynes claimed originality for the symbol in correspondence with W. H. Macaulay in 1907 (Aldrich, 2020). In his book, he says that had not been aware of McColl's earlier notation when he devised the symbol (Keynes, 1921, p. 177). In a review of Keynes's book, Broad suggested that Keynes had borrowed the symbol from Johnson (Broad, 1922, p. 78), but Johnson acknowledged Keynes's priority, at least in publication. Johnson read Keynes's dissertation and likely used Keynes's symbol subsequently in lectures attended by Broad and Dorothy Wrinch (Aldrich, 2008, 2020).

a probability measure (p. 11), Estes and Suppes explained that "the experimenter may conditionalize the probabilities of reinforcement upon preceding events of the sample space in whatever manner he pleases" (pp. 20–21). The use of "update" in this context seems to have appeared much later, only in the late 1970s.

In the 1960s, A. P. Dempster was writing about his own rules for or of combination and conditioning and comparing them with Bayesian rules (Dempster, 1967, 1968). In my 1976 book on the Dempster–Shafer theory (Shafer, 1976), I distinguished between *Bayes's rule of conditioning*, as I called it, and Bayes's theorem.

- Bayes's rule of conditioning says that when you learn A, you change your probability for B from $P(B)$ to $P(B|A)$ as given by (1), regardless of the order in which the events may have happened in the world. I attributed this rule directly to Bayes because he had given a betting argument for it, which is erroneous in my opinion; see Shafer (1982).
- Bayes's theorem is more specific; it is the Bayesian rule for changing probabilities for a parameter based on observations (or, in Laplace's words, obtaining probabilities for causes from events). Beginning with Cournot843 (1843), some authors called this Bayes's rule (*règle de Bayes* in French; *Bayesschen Regel* in German); others called it Bayes's formula or Bayes's theorem.[5] In English, it was often called the method of "inverse probability." Now that (1) is regarded as a definition, it is more often called a theorem.

The distinction between Bayes's rule of conditioning (or updating or conditionalization; see Teller (1973)) and Bayes's theorem is now widely made, but it is remains unfamiliar to many statisticians. Perhaps for this reason, Gong and Meng blur the distinction, calling

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)}$$

"Bayes rule." I find this confusing, because when (1) is treated as a general rule for updating a probability measure after observing an event, there is no presumption that the probabilities of the event conditional on all other events had previously been singled out and calculated.

6. THE IMPLICATIONS OF INSISTING ON A PROTOCOL

Gong and Meng are kind enough to cite the 1985 article in which I insisted that Bayesian updating after learning B is legitimate only in the presence of a protocol that singled out B as one of the things we might learn (Shafer, 1985). It is only in this case, I argued that De Moivre's betting argument and its variants (e.g., de Finetti, 1937, Teller, 1973) justify Bayes's rule of conditioning and only in this case that paradox can be avoided. I would like to add to their discussion an explanation of how I understand the consequences of insisting on a protocol.

By a *protocol*, I mean what Joseph L. Doob and later probabilists have called a *filtration*. Starting at time 0, you first learn X_1, then X_2, etc. In the simple special case where these variables are all binary and we stop at fixed time n, we can visualize the protocol as a binary tree. The sample space Ω is the set of all paths through the tree, from time 0 to time n. There are 2^n elements in Ω and hence 2^{2^n} events. But there are exponentially fewer nodes in the tree—only $2^n - 1$. But only a node in the tree can represent what you may have learned at some point in time. If and when you reach a particular node, say by observing $x_1, \ldots, x_k$, your new probability for an event A will be your original probability "conditioned" on $X_1 = x_1, \ldots, X_k = x_k$. But you will never "condition" on any of the $2^{2^n} - 2^n + 1$ events not of this form. So the notion that you have a methodology that allows you to "update" when your new information is any subset B of Ω is illusory.

A common Bayesian response is that you should of course condition on everything you have learned, including the fact that you learned it. This implies that the elements of Ω specify what you will and will not learn at every point in time. So the Bayesian view already implicitly calls for a protocol for how new information may arrive. In my view, leaving this need for a protocol implicit is more than an invitation to paradox. It is deceptive. Once the demand to provide a probability model for your entire learning process is made explicit, it becomes obvious that the demand often cannot be satisfied.

Surely we should conclude that models with updating rules are only one limited set of tools for assessing uncertainty. We also need ideas for evaluating and combining unanticipated evidence, such as Jacob Bernoulli proposed in (Bernoulli, 1713, 2006, Part IV, Chapter 3), Dempster and I proposed in the 1960s and 1970s, and others have proposed before and since.

ACKNOWLEDGMENTS

My preparation of this note has benefited from recent conversations with John Aldrich, Bernard Bru, Roubin Gong, Xiao-Li Meng and Sandy Zabell, and from countless conversations over the years with other colleagues.

[5] Bayes's friend and executor, Richard Price, used the phrase "Mr. Bayes's rules" to refer to formulas Bayes had derived for approximating what we now call posterior and predictive Bayesian probabilities in the binomial problem (Dale, 1999, pp. 39–40).

REFERENCES

ALDRICH, J. (2008). Keynes among the statisticians. *Hist. Polit. Econ.* **40** 265–316.

ALDRICH, J. (2020). Personal communication.
BERNOULLI, J. (1713). *Ars Conjectandi*. Thurnisius, Basel. See Bernoulli (2006), translation by Edith Sylla.
BERNOULLI, J. (2006). *The Art of Conjecturing. Together with "Letter to a Friend on Sets in Court Tennis*. Johns Hopkins Univ. Press, Baltimore, MD. MR2195221
BERTRAND, J. (1889). *Calcul des Probabilités*. Gauthier-Villars, Paris.
BOOLE, G. (1854). *An Investigation of the Laws of Thought, on Which Are Founded the Mathematical Theories of Logic and Probabilities*. Macmillan, London. Reprinted by Dover, New York, 1958.
BROAD, C. D. (1914). *Perception, Physics and Reality: An Enquiry into the Information That Physical Science Can Supply About the Real*. Cambridge Univ. Press, Cambridge.
BROAD, C. D. (1922). *A treatise on probability*. By J. M. Keynes. *Mind* **31** 72–85.
BRU, B. (1989). Doutes de d'Alembert sur le calcul des probabilités. In *Jean D'Alembert, Savant et Philosophe. Portrait à Plusieurs Voix* (M. Emery and P. Monzani, eds.) 279–292. Archives contemporaines, Paris.
BRU, B. (2002). Des fraises et des oranges. In *Sciences, Musiques, Lumières: Mélanges Offerts à Anne-Marie Chouillet* (U. Kölving and I. Passeron, eds.) 3–10. Centre international d'étude du XVIII^{e} siècle, Ferny-Voltaire.
COURNOT, A. A. (1843). *Exposition de la Théorie des Chances et des Probabilités*. Hachette, Paris. Reprinted in 1984 as Volume I (Bernard Bru, editor) of Cournot (1973–2010).
COURNOT, A. A. (1973–2010). *Œuvres Complètes*. Vrin, Paris. The volumes are numbered I through XI, but VI and XI are double volumes.
CZUBER, E. (1908). *Wahrscheinlichkeitsrechnung und Ihre Anwendung Auf Fehlerausgleichung, Statistik und Lebensversicherung* **1**, 2nd ed. Teubner, Leipzig.
D'ALEMBERT, J. L. R. (1767). *Éclaircissmens sur Différens Endroits des* Élémens *de Philosophie*. Chatelain, Amsterdam. In the 5th volume of d'Alembert's *Mélanges de littéraire d'histoire et de philosophie*.
DALE, A. I. (1999). *A History of Inverse Probability from Thomas Bayes to Karl Pearson*, 2nd ed. *Sources and Studies in the History of Mathematics and Physical Sciences*. Springer, New York. MR1710390 https://doi.org/10.1007/978-1-4419-8652-8
DE FINETTI, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré* **7** 1–68. MR1508036
DEMPSTER, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38** 325–339. MR0207001 https://doi.org/10.1214/aoms/1177698950
DEMPSTER, A. P. (1968). A generalization of Bayesian inference. (With discussion). *J. Roy. Statist. Soc. Ser. B* **30** 205–247. MR0238428
DE MOIVRE, A. (1738). *The Doctrine of Chances: Or, a Method of Calculating the Probabilities of Events in Play*, 2nd ed. Pearson, London.
ESTES, W. K. and SUPPES, P. (1957). Foundations of Statistical Learning Theory, I. The linear model for simple learning. Technical Report No. 16. Behavioral Sciences Division, Applied Mathematics and Statistics Laboratory, Stanford Univ.
FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh. Subsequent editions appeared in 1959 and 1973.
GORROOCHURN, P. (2012). *Classic Problems of Probability*. Wiley, Hoboken, NJ. MR2976816 https://doi.org/10.1002/9781118314340
HAUSDORFF, F. (1901). Beiträge zur Wahrscheinlichkeitsrechnung. *Sitzungsberichte der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse* **53** 152–178.
JEFFREYS, H. (1931). *Scientific Inference*, 1st ed. Cambridge Univ. Press, Cambridge.
JOHNSON, W. E. (1924). *Logic* **3**. Cambridge Univ. Press, Cambridge.
KEYNES, J. M. (1921). *A Treatise on Probability*. Macmillan, London.
KNEBEL, S. K. (2000). *Wille, Würfel und Wahrscheinlichkeit: Das System der Moralischen Notwendigkeit in der Jesuitenscholastik* 1550–1700. Meiner, Hamburg.
KOLMOGOROV, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin. English translation *Foundations of the Theory of Probability*, Chelsea, New York, 1950, 2nd ed. 1956.
LACROIX, S. F. (1816). *Traité Élémentaire du Calcul des Probabilités*. Courcier, Paris. Second edition 1822.
LIAGRE, J.-B.-J. (1852). *Calcul des Probabilités et Théorie des Erreurs Avec des Applications aux Sciences D'observation en Général et à la Géodésie*. Muquardt, Brussels. Second edition, 1879, prepared with the assistance of Camille Peny.
MACCOLL, H. (1896/97). The Calculus of Equivalent Statements. (Sixth Paper.). *Proc. Lond. Math. Soc.* **28** 555–579. MR1576659 https://doi.org/10.1112/plms/s1-28.1.555
MARKOV, A. A. (1900). *Probability Calculus* (*in Russian*). Imperial Academy, St. Petersburg. The second edition, which appeared in 1908, was translated into German as *Wahrscheinlichkeitsrechnung*, Teubner, Leipzig, Germany, 1912.
MCCOLL, H. (1879/80). The Calculus of Equivalent Statements. (Fourth Paper.). *Proc. Lond. Math. Soc.* **11** 113–121. MR1575250 https://doi.org/10.1112/plms/s1-11.1.113
MCCOLL, H. (1880/81). A Note on Prof. C. S. Peirce's Probability Notation of 1867. *Proc. Lond. Math. Soc.* **12** 102. MR1575784 https://doi.org/10.1112/plms/s1-12.1.102
MIROWSKI, P., ed. (1994). *Edgeworth on Chance, Economic Hazard, and Statistics*. Rowman & Littlefield, Lanham, MD.
PEIRCE, C. S. (1867). On an improvement in Boole's calculus of logic. *Proc. Amer. Acad. Arts Sci.* **7** 250–261.
SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton, NJ. MR0464340
SHAFER, G. (1982). Bayes's two arguments for the rule of conditioning. *Ann. Statist.* **10** 1075–1089. MR0673644
SHAFER, G. (1985). Conditional probability. *Int. Stat. Rev.* **53** 261–277.
SHAFER, G. and VOVK, V. (2006). The sources of Kolmogorov's *Grundbegriffe*. *Statist. Sci.* **21** 70–98. MR2275967 https://doi.org/10.1214/088342305000000467
TELLER, P. (1973). Conditionalization and observation. *Synthese* **26** 218–258.
WRINCH, D. and JEFFREYS, H. (1919). On some aspects of the theory of probability. *The London, Edinburgh and Dublin Philosophical Magazine, Series* 6 **38** 715–731.