The Boxer, the Wrestler, and the Coin Flip: A Paradox of Robust Bayesian Inference and Belief Functions

Andrew GELMAN

Bayesian inference requires all unknowns to be represented by probability distributions, which awkwardly implies that the probability of an event for which we are completely ignorant (e.g., that the world's greatest boxer would defeat the world's greatest wrestler) must be assigned a particular numerical value such as 1/2, as if it were known as precisely as the probability of a truly random event (e.g., a coin flip).

Robust Bayes and *belief functions* are two methods that have been proposed to distinguish ignorance and randomness. In robust Bayes, a parameter can be restricted to a range, but without a prior distribution, yielding a range of potential posterior inferences. In belief functions (also known as the Dempster-Shafer theory), probability mass can be assigned to subsets of parameter space, so that randomness is represented by the probability distribution and uncertainty is represented by large subsets, within which the model does not attempt to assign probabilities.

Through a simple example involving a coin flip and a boxing/wrestling match, we illustrate difficulties with robust Bayes and belief functions. In short: robust Bayes allows ignorance to spread too broadly, and belief functions inappropriately collapse to simple Bayesian models.

KEY WORDS: Dempster-Shafer theory; Epistemic and aleatory uncertainty; Foundations of probability; Ignorance; Robust Bayes; Subjective prior distribution.

1. USING PROBABILITY TO MODEL BOTH RANDOMNESS AND UNCERTAINTY

We define two binary random variables: the outcome X of a coin flip, and the outcome Y of a hypothetical fight to the death between the world's greatest boxer and the world's greatest wrestler (Figure 1):

$$X = \begin{cases} 1 & \text{if the coin lands "heads"} \\ 0 & \text{if the coin lands "tails"} \end{cases}$$

$$Y = \begin{cases} 1 & \text{if the boxer wins} \\ 0 & \text{if the wrestler wins.} \end{cases}$$

In the Bayesian framework, X and Y must be given probability distributions. Modeling X is easy: $\Pr(X = 1) = \Pr(X = 0) = 1/2$, probabilities that can be justified on physical grounds. [The outcomes of a coin caught in mid-air can be reasonably modeled as equiprobable (see, e.g., Jaynes 1996; Gelman and Nolan 2002) but if this makes you uncomfortable, you can think of X as being defined based on a more purely random process such as a radiation counter.]

Modeling Y is more of a challenge, because we have little information to directly bear on the problem and (let us suppose) no particular reason for favoring the boxer or the wrestler in the bout. We shall consider this a problem of *ignorance*, the modeling of which has challenged Bayesians for centuries and, indeed, has no clearly defined solution (hence the jumble of "noninformative priors" and "reference priors" in the statistical literature). The distinction between X and Y is between randomness and ignorance or, as characterized by O'Hagan (2004), between aleatory and epistemic uncertainty.

Here we will model Y as a Bernoulli random variable, with $\Pr(Y = 1) = \pi$, and assign a uniform prior distribution to π on the range $[0, 1]$. That is, we assume complete ignorance about

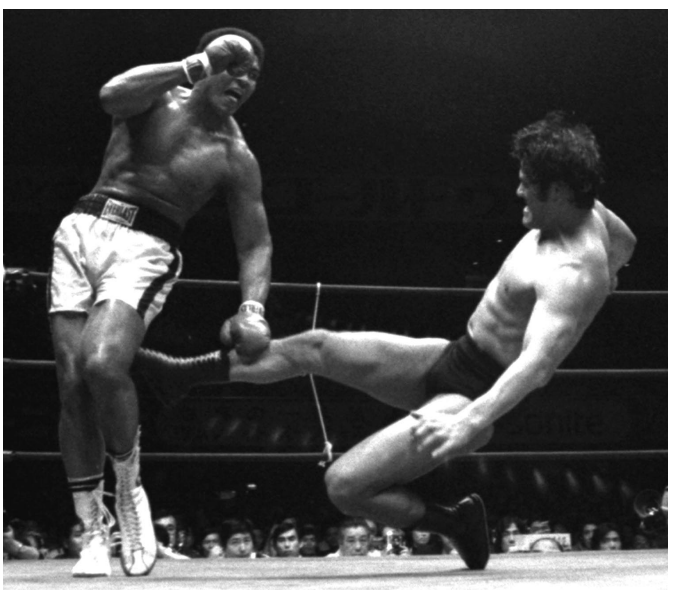

Figure 1. Wrestler Antonio Inoki delivers a flying kick to Muhammad Ali during their exhibition on June 26, 1976. Used with permission from the Stars and Stripes. Copyright 1976, 2006 Stars and Stripes.

Andrew Gelman is Professor, Department of Statistics and Department of Political Science, Columbia University, 1016 Social Work Building, New York, NY 10027 (E-mail: *gelman@stat.columbia.edu*; Web: *www.stat.columbia.edu/∼gelman*). The author thanks Arthur Dempster, Augustine Kong, Hal Stern, David Krantz, Glenn Shafer, Larry Wasserman, Jouni Kerman, and an anonymous reviewer for helpful comments, and the National Science Foundation for financial support.

  DOI: 10.1198/000313006X106190

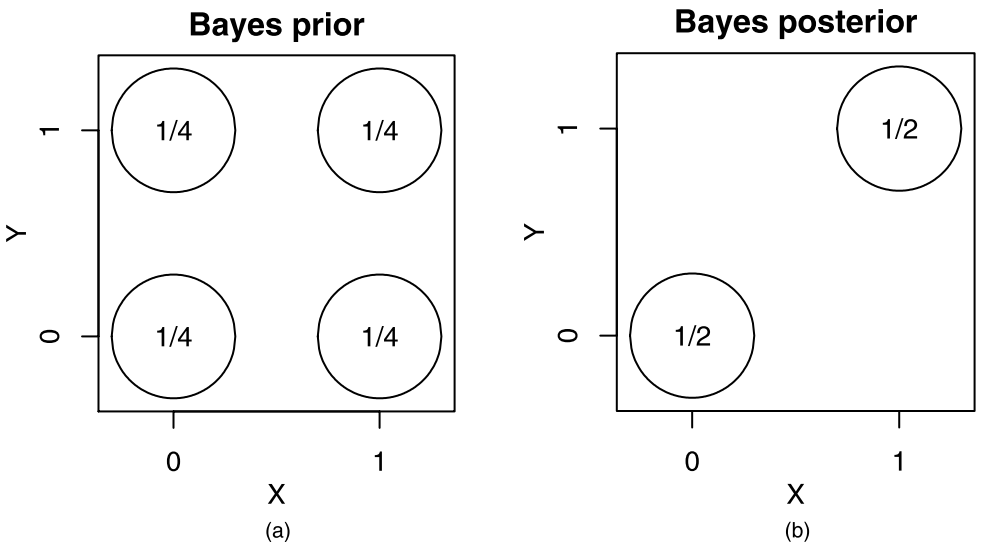

Figure 2. (a) Prior predictive distribution and (b) posterior predictive distribution (after conditioning on X = Y) for the coin-flip/fighting-match example, under pure Bayesian inference. X, which we understand completely, and Y, for which we have complete ignorance, are treated symmetrically.

the probability that the boxer wins. [As reviewed by Bernardo and Smith (1994) and Kass and Wasserman (1996), a uniform distribution on the probability scale is only one of the many ways to define "ignorance" in this sort of problem. Our key assumption here is symmetry: that we have no particular reason to believe either the boxer or the wrestler is superior.]

Finally, the joint distribution of X and Y must be specified. We shall assume the coin flip is performed apart from and with no connection to the boxing/wrestling match, so that it is reasonable to model the two random variables as independent.

2. A SIMPLE EXAMPLE OF CONDITIONAL INFERENCE, FROM GENERALIZED BAYESIAN PERSPECTIVES

As in the films *Rashomon* and *The Aristocrats*, we shall tell a single story from several different perspectives. The story is as follows: X and Y are defined above, and we now learn that $X = Y$: either the coin landed heads and the boxer won, or the coin landed tails and the wrestler won. To clarify the information available here: we suppose that a friend has observed the fight and the coin flip and has agreed ahead of time to tell us if $X = Y$ or $X \neq Y$. It is thus appropriate to condition on the event $X = Y$ in our inference.

Conditioning on $X = Y$ would seem to tell us nothing useful—merely the coupling of a purely random event to a purely uncertain event—but, as we shall see, this conditioning leads to different implications under different modes of statistical inference.

In straight Bayesian inference the problem is simple. First off, we can integrate the parameter π out of the distribution for Y to obtain $\Pr(Y = 1) = \Pr(Y = 1) = 1/2$. Thus, X and Y—the coin flip and the fight—are treated identically in the probability model, which we display in Figure 2. In the prior distribution, all four possibilities of X and Y are equally likely; after conditioning on $X = Y$, the two remaining possibilities are equally likely. (We label the plots in Figures 2 and 3 as *predictive* distributions because they show the probabilities of observables rather than parameters.)

There is nothing wrong with this inference, but we might feel uncomfortable giving the model for the uncertain Y the same inferential status as the model for the random X. This is a fundamental objection to Bayesian inference—that complete ignorance is treated mathematically the same as an event with probabilities known from physical principles. The distinction between randomness and ignorance has been addressed using robust Bayes and belief functions.

2.1 Robust Bayes

Robust Bayes is a generalization of Bayesian inference in which certain parameters are allowed to fall in a range but without being specified a prior distribution. Or, to put it another way, a continuous range of models is considered, yielding a continuous range of possible posterior inferences (Berger 1984, 1990; Wasserman 1992).

For our example, we can use robust Bayes to model complete ignorance by allowing π—the probability that Y equals 1, that the boxer defeats the wrestler—to fall anywhere in the range $[0, 1]$. Figure 3(a) displays the prior distribution, and Figure 3(b) displays the posterior distribution after conditioning on the event $X = Y$.

Because we are allowing the parameter π to fall anywhere between 0 and 1, the robust Bayes inference leaves us with complete uncertainty about the two possibilities $X = Y = 0$ and $X = Y = 1$. This seems wrong in that it has completely degraded our inferences about the coin flip, X. Equating it with an event we know nothing about—the boxing/wrestling match—has led us to the claim that we can say nothing at all about the coin flip. It would seem more reasonable to still allow a 50/50 probability for X—but this cannot be done in the robust Bayes framework in which the entire range of π is being considered. [More precise inferences would be obtained by restricting π to a narrower range such as $[0.4, 0.6]$, but in this example we specifically want to model complete ignorance.] This is an issue

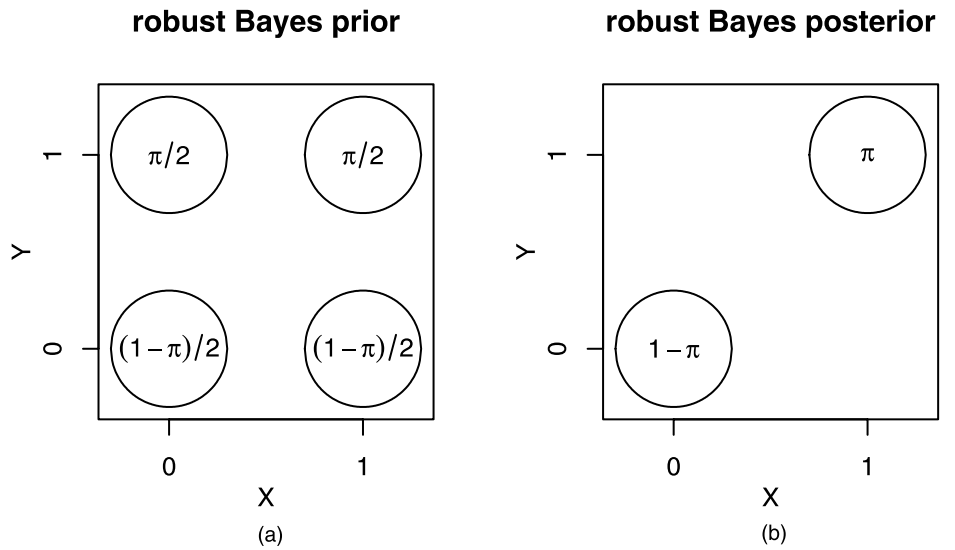

Figure 3. (a) Prior predictive distribution and (b) posterior predictive distribution (after conditioning on X = Y) for the coin-flip/fighting-match example, under robust Bayesian inference in which the parameter π is allowed to fall anywhere in the range [0, 1]. After conditioning on X = Y, we can now say nothing at all about X, the outcome of the coin flip.

that inevitably arises when considering ranges of estimates (e.g., Imbens and Manski 2004), and it is not meant to imply that robust Bayes is irredeemably flawed, but rather to indicate a counterintuitive outcome of using the range $\pi \in [0, 1]$ to represent complete ignorance. Seidenfeld and Wasserman (1993) showed that this "dilation" phenomenon—conditional inferences that are less precise than marginal inferences—is inevitable in robust Bayes. By modeling π with complete ignorance, we have constructed an extreme example of dilation.

2.2 Belief Functions

The method of *belief functions* (Dempster 1967, 1968) has been proposed as a generalization of Bayesian inference that more directly allows the modeling of ignorance (Shafer 1976). In belief functions, probability mass can be assigned to arbitrary subsets of the sample space—thus generalizing Bayesian inference, which assigns probability mass to atomic elements of the space.

We briefly review belief functions and then apply them to our example. For a binary variable, the probability mass of a belief function can be distributed over all nonempty subsets of the sample space: $\{0\}$, $\{1\}$, and $\{0, 1\}$. For example, a coin flip would be assigned probability masses $p(\{0\}) = 0.5$, $p(\{1\}) = 0.5$, $p(\{0, 1\}) = 0$; and a random outcome with probability p of success would be assigned probability masses $p(\{0\}) = 1 - p$, $p(\{1\}) = p$, $p(\{0, 1\}) = 0$. These are simply Bayesian probability assignments.

Belief functions become more interesting when used to capture uncertainty. For example, consider a random outcome with probability p of success, with p itself known to fall somewhere between a *lower probability* of 0.4 and an *upper probability* of 0.9. In belief functions, the lower probability of a set A is defined as the sum of the probability masses assigned to subsets of A (including A itself), and the upper probability is the sum of the probability masses of all sets that intersect with A. For a binary variable, this definition just reduces to: $\Pr(0) \in [p(\{0\}),\ p(\{0\}) + p(\{0, 1\})]$ and $\Pr(1) \in [p(\{1\}),\ p(\{1\}) + p(\{0, 1\})]$. This can be represented by a belief function with probability masses $p(\{0\}) = 0.1$, $p(\{1\}) = 0.4$, $p(\{0, 1\}) = 0.5$. In this model, the probability of the event "0" is somewhere between 0.1 and 0.6, and the probability of the event "1" is somewhere between 0.4 and 0.9.

Statistical analysis is performed by expressing each piece of available information as a belief function over the space of all unknowns, then combining them using "Dempster's rule," a procedure which we do not present in general here but will illustrate for our simple problem. Dempster's rule differs from the robust Bayes approach described earlier in combining the underlying probability masses of the belief functions, not the upper and lower probablities which are computed only at the end of the analysis.

Belief functions can be applied to the boxer/wrestler problem in two steps. First, X is given a straight probability distribution, just as in Bayesian inference, with 50% probability on each outcome. Second, Y is given a so-called vacuous belief function, assigning 100% of the probability mass to the set $\{0,1\}$, thus stating our complete ignorance in the outcome of the fight. The events X and Y would still be independent, and their joint belief function is shown in Figure 4(a)—it has two components, each assigned belief 0.5.

Conditioning on $X = Y$ (i.e., combining with the belief function that assigns 100% of its probability mass to the set $\{(0, 0), (1, 1)\}$) yields the belief function shown in Figure 4(b). Oddly enough, all the vacuity has disappeared and the resulting inference is identical to the pure Bayes posterior distribution in Figure 2(b). This does not seem right at all: coupling the fight outcome Y to a purely random X has caused the belief function for Y to collapse from pure ignorance to a simple 50/50 probability distribution. No information has been added, yet the belief function has changed dramatically. Once again, we would not use this example to dismiss belief functions [see Shafer (1990) for some background on their theory and application], but this example does suggest that the belief-function modeling of ignorance is potentially fragile.

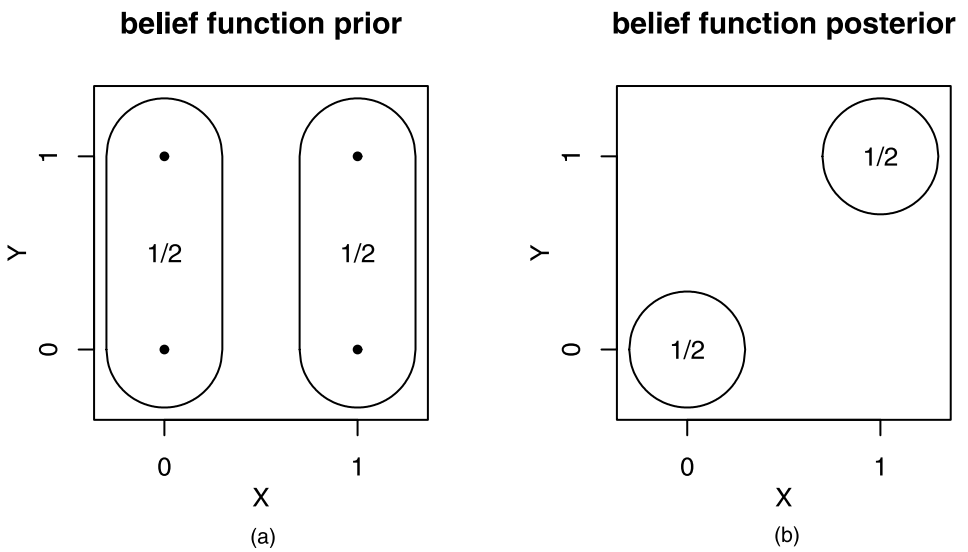

Figure 4. (a) Prior belief function and (b) posterior belief function (after conditioning on X = Y) for the coin-flip/fighting-match example, under Dempster-Shafer inference in which 100% of the prior mass for Y is placed on the set {0, 1}. After combining with the information that X = Y, the belief function for Y has collapsed to the simple case of equal probabilities.

When shown this example, Arthur Dempster, the inventor of belief functions (Dempster 1967, 1968), commented that a step of our analysis is combining the prior information with the information $Y = X$. He wrote,

> The Dempster-Shafer information fusion rule is a common sense combination of Bayesian conditioning and Boolean logic. What it depends on crucially, including in its two special cases, is independence. Independence can only be assessed in a realistic situation, preferably a specific real situation, and always needs careful situation-specific assessment. When I try to create a detailed realistic story around your hypothetical example, I come out reasonably comfortable about the belief-function posterior (0.5, 0.5, 0). There is relevant information in the observer's report, and this justifies the change from "don't know" to precise probabilities. Hence I feel that the general framework is supported.

In a separate communication, Glenn Shafer, the other key developer of belief functions (Shafer 1976, 1990), also pointed to the artificiality of the $X = Y$ setup and wrote,

> I am not quite with you when you declare your various results are counterintuitive. Because the example is so artificial, I don't see a source for intuition. . . . as always, the question is whether the "independence" holds for the rule to be appropriate in the particular case.

I recognize these concerns, and I agree that the ultimate test of a statistical method is in its applications, but I am still disturbed by this particular automatic application of belief functions.

3. DISCUSSION

Bayesian inference is an extremely powerful tool in applied statistics (see, e.g., Carlin and Louis 2001; Gelman, Carlin, Stern, and Rubin 2003), but an ongoing sticking point is the necessity for prior distributions, which are particularly controversial when used to model ignorance. [Prior probabilities and Bayesian inference can also be motivated as necessary for coherent decision making (Keynes 1921; Cox 1925; von Neumann and Morgenstern 1944) but this just shifts the problem to a requirement of coherent decision making under ignorance, which in practice might be no easier than assigning prior probabilities directly.] Various generalizations of Bayesian inference, including robust Bayes and belief functions, have been proposed to ease this difficulty by mathematically distinguishing between uncertainty and randomness. Using a simple example coupling a completely known probability (for a coin flip) with a completely unknown probability (for the fight), we have shown that robust Bayes and belief functions can yield counterintuitive results. We conclude that the challenge of assigning prior distributions is real, and we do not see any easy way of separating uncertainty from probability. However, we have not considered other forms of inference such as fuzzy logic (Zadeh 1965), which can perhaps resolve these problems, at least for some categories of examples.

Another approach is the frequentist or randomization approach to inference, under which probability can only be assigned to random events (i.e., those defined based on a physical randomization process with known probabilities) and never to uncertainties, which must be represented purely by unmodeled parameters (see, e.g., Cox and Hinkley 1974.) For our example, Y will not be assigned a probability distribution at all, and so the operation of conditioning on $X = Y$ cannot be interpreted probabilistically, and no paradox arises. The difficulty of frequentist inference is its conceptual rigidity—taking its prescriptions literally, one would not be allowed to model business forecasts, industrial processes, demographic patterns, or for that matter real-life sample surveys, all of which involve uncertainties that cannot be simply represented by physical randomization. [Jaynes (1996) and Gelman et al. (2003, chap. 1) discussed various examples of probability models that are empirically-defined but do not directly correspond to long-run frequencies.] Our point here is not to debate Bayesian versus frequentist notions of probability but rather to note that the difficulty of modeling both uncertainty and randomness is tied to the flexibility of Bayesian modeling.

Finally, how can the distinction between uncertainty and randomness be understood in Bayesian theory? O'Hagan (2004) provided a clear explanation, comparing a coin flip to an equivalent of our boxer/wrestler example. In Bayesian inference, our prior predictive distributions for X and for Y are identical, which does not seem quite right, since we understand the pro-

cess generating X so much better than that of Y. The difficulty is that the integral of a probability is a probability: for the model of Y, integrating out the uncertainty in π simply yields $\Pr(Y = 1) = \Pr(Y = 0) = 1/2$.

As discussed by O'Hagan, the resolution of the paradox is that probabilities, and decisions, do not take place in a vacuum. If the only goal were to make a statement, or a bet, about the outcome of the coin flip or the boxing/wrestling match, then yes, $p = 1/2$ is all that can be said. But the events occur within a context. In particular, the coin flip probability remains at 1/2, pretty much no matter what information is provided (before the actual flipping occurs, of course). In the coin-flipping example, one can reframe the model as the probability of heads having a point-mass prior at 0.5—in some sense, this is the best-case prior information about the probability before the coin is flipped.

In contrast, one could imagine gathering lots of information (e.g., reports of previous fights such as the exhibition between Antonio Inoki and Muhammad Ali pictured in Figure 1) that would refine one's beliefs about π. [Actually, the Ali–Inoki match was said by many wrestling experts to be a show rather than a serious competitive fight; see, e.g., Draeger (2000), but for the purposes of this argument we shall consider it as representative of the sort of information that could be used in constructing an informative prior distribution.] Averaging over uncertainty in π, the probability the boxer wins is $\Pr(Y = 1) = \mathrm{E}(\pi)$, which equals 1/2 for a uniform prior distribution on π but can change as information is gathered about π. Uncertainty in π (in O'Hagan's terms, "epistemic uncertainty") necessarily maps to potential information we could learn that would tell us something about π. So in this larger, potentially hierarchical, context, Bayesian inference can distinguish between aleatory uncertainty (randomness) and epistemic uncertainty (ignorance).

Such an approach does not eliminate the difficulties of using probability to model uncertainty—in particular, "noninformative" or similarly weak prior distributions still must be chosen in some way (Kass and Wasserman 1996) but it can limit the damage resulting from an inappropriate choice of prior.

[Received September 2005. Revised November 2005.]

REFERENCES

Berger, J. O. (1984), "The Robust Bayesian Viewpoint" (with discussion), in *Robustness in Bayesian Statistics*, ed. J. Kadane, Amsterdam: North-Holland.

——— (1990), "Robust Bayesian Analysis: Sensitivity to the Prior," *Journal of Statistical Planning and Inference*, 25, 303–328.

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Carlin, B. P., and Louis, T. A. (2001), *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.), London: Chapman and Hall.

Cox, R. T. (1925), *The Algebra of Probable Inference*, Baltimore, MD: Johns Hopkins University Press.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Dempster, A. P. (1967), "Upper and Lower Probabilities Induced by a Multivalued Mapping," *Annals of Mathematical Statistics*, 38, 205–247.

Dempster, A. P. (1968), "A Generalization of Bayesian Inference," *Journal of the Royal Statistical Society*, Series B, 30, 205–247.

Draeger, D. F. (2000), "Muhammed Ali versus Antonio Inoki," *Journal of Combative Sport*, January, *ejmas.com/jcs/jcsdraeger_alivsinoki.htm*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis* (2nd ed.), London: Chapman and Hall.

Gelman, A., and Nolan, D. (2002), "You Can Load a Die but You Can't Bias a Coin Flip," *The American Statistician*, 56, 308–311.

Imbens, G., and Manski, C. (2004), "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

Jaynes, E. T. (1996), *Probability Theory: The Logic of Science*. Available online at *bayes.wustl.edu/etj/prob.html*.

Kass, R. E., and Wasserman, L. (1996), "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association*, 91, 1343–1370.

Keynes, J. M. (1921), *A Treatise on Probability*, London: Macmillan.

O'Hagan, A. (2004), "Dicing With the Unknown," *Significance*, 1, 132–133.

Seidenfeld, T., and Wasserman, L. (1993), "Dilation for Sets of Probabilities," *The Annals of Statistics*, 21, 1139–1154.

Shafer, G. (1976), *A Mathematical Theory of Evidence*, Princeton, NJ: Princeton University Press.

Shafer, G. (1990), "Perspectives on the Theory and Practice of Belief Functions," *International Journal of Approximate Reasoning*, 4, 323–362.

von Neumann, J., and Morgenstern, O. (1944), *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press.

Wasserman, L. (1992), "Recent Methodological Advances in Robust Bayesian Inference" (with discussion), in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Cambridge, MA: Oxford University Press, pp. 438–502.

Zadeh, L. A. (1965), "Fuzzy Sets," *Information and Control*, 8, 338–353.