Statistical Inference: The Big Picture

Robert E. Kass[*]

December 1, 2009

Abstract

Statistics has moved beyond the frequentist-Bayesian controversies of the past. Where does this leave our ability to interpret results? I suggest that a philosophy compatible with statistical practice, labelled here *statistical pragmatism*, serves as a foundation for inference. Statistical pragmatism is inclusive and emphasizes the assumptions that connect statistical models with observed data. I argue that introductory courses often mis-characterize the process of statistical inference and I propose an alternative "big picture" depiction.

Keywords: Bayesian, confidence, frequentist, statistical education, statistical pragmatism, statistical significance.

[*] Robert E. Kass is Professor, Department of Statistics, Center for the Neural Basis of Cognition, and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, 15217 (email: kass@stat.cmu.edu). This work was supported in part by NIH grant MH064537. The author is grateful for comments on an earlier draft by Brian Junker, Nancy Reid, Steven Stigler, Larry Wasserman, and Gordon Weinberg.

1 INTRODUCTION

The protracted battle for the foundations of statistics, joined vociferously by Fisher, Jeffreys, Neyman, Savage, and many disciples, has been deeply illuminating, but it has left statistics without a philosophy that matches contemporary attitudes. Because each camp took as its goal exclusive ownership of inference, each was doomed to failure. We have all, or nearly all, moved past these old debates, yet, our textbook explanations have not caught up with the eclecticism of statistical practice.

The difficulties go both ways. Bayesians have denied the utility of confidence and statistical significance, attempting to sweep aside the obvious success of these concepts in applied work. Meanwhile, for their part, frequentists have ignored the possibility of inference about unique events despite their ubiquitous occurrence throughout science. Furthermore, intepretations of posterior probability in terms of subjective belief, or confidence in terms of long-run frequency, give students a limited and sometimes confusing view of the nature of statistical inference. When used to introduce the expression of uncertainty based on a random sample, these caricatures forfeit an opportunity to articulate a fundamental attitude of statistical practice.

Most modern practitioners have, I think, an open-minded view about alternative modes of inference, but are accutely aware of theoretical assumptions and the many ways they may be mistaken. I would suggest that it makes more sense to place in the center of our logical framework the match or mis-match of theoretical assumptions with the real world of data. This, it seems to me, is the common ground that Bayesian and frequentist statistics share; it is more fundamental than either paradigm taken separately; and as we strive to foster widespread understanding of statistical reasoning, it is more important for beginning students to appreciate the role of theoretical assumptions than for them to recite correctly the long-run interpretation of confidence intervals. In the hope of prodding our discipline to right a lingering imbalance, I attempt here to describe the dominant contemporary philosophy of statistics.

2 STATISTICAL PRAGMATISM

I propose to call this modern philosophy *statistical pragmatism.* I think it is based on the following attitudes:

1. Confidence, statistical significance, and posterior probability are all valuable inferential tools.

2. Simple chance situations, where counting arguments may be based on symmetries that generate equally-likely outcomes (6 faces on a fair die; 52 cards in a shuffled deck), supply basic intuitions about probability. Probability may be built up to important but less immediately intuitive situations using abstract mathematics, much the way real numbers are defined abstractly based on intuitions coming from fractions. Probability is usefully calibrated in terms of fair bets: another way to say the probability of rolling a 3 with a fair die is 1/6 is that 5 to 1 odds against rolling a 3 would be a fair bet.

3. Long-run frequencies are important mathematically, interpretively, and pedagogically. However, it is possible to assign probabilities to unique events, including rolling a 3 with a fair die or having a confidence interval cover the true mean, without considering long-run frequency. Long-run frequencies may be regarded as consequences of the law of large numbers rather than as part of the definition of probability or confidence.

4. Similarly, the subjective interpretation of posterior probability is important as a way of understanding Bayesian inference, but it is not fundamental to its use: in reporting a 95% posterior interval one need not make a statement such as, “My personal probability of this interval covering the mean is .95.”

5. Statistical inferences of all kinds use statistical models, which embody theoretical assumptions. As illustrated in Figure 1, like scientific models, statistical models exist in an abstract framework; to distinguish this framework from the real world inhabited by data we may call it a “theoretical world.” Random variables, confidence intervals, and posterior probabilities all live in this theoretical world. When we use a

> statistical model to make a statistical inference we implicitly assert that the variation exhibited by data is captured reasonably well by the statistical model, so that the theoretical world corresponds reasonably well to the real world. Conclusions are drawn by applying a statistical inference technique, which is a theoretical construct, to some real data. Figure 1 depicts the conclusions as straddling the theoretical and real worlds. Statistical inferences may have implications for the real world of new observable phenomena, but in scientific contexts, conclusions most often concern scientific models (or theories), so that their "real world" implications (involving new data) are somewhat indirect (the new data will involve new, and different experiments).

The statistical models in Figure 1 could involve large function spaces or other relatively weak probabilistic assumptions. Careful consideration of the connection between models and data is a core component of both the art of statistical practice and the science of statistical methodology. The purpose of Figure 1 is to shift the grounds for discussion.

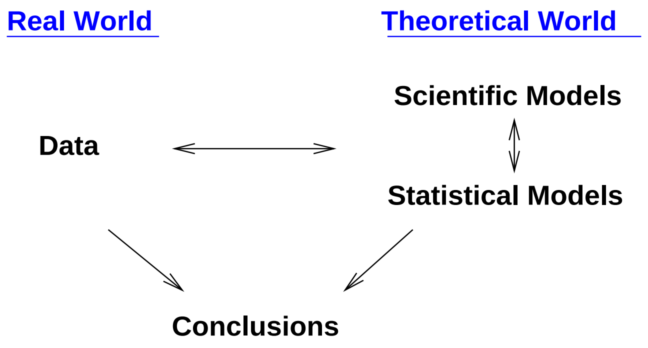

Figure 1: *The big picture of statistical inference. Statistical procedures are abstractly defined in terms of mathematics but are used, in conjunction with scientific models and methods, to explain observable phenomena.*

Note, in particular, that data should not be confused with random variables. Random variables live in the theoretical world. When we say things like, "Let us assume the data are normally distributed" and we proceed to make a statistical inference, we do not need to take these words literally as asserting that the data form a random sample. Instead, this kind of language is a convenient and familiar shorthand for the much weaker assertion that, for our specified purposes, the variability of the data are adequately consistent with variability that would occur in a random sample. This linguistic amenity is used routinely in both frequentist and Bayesian frameworks. Historically, the distinction between data and random variables, the match of the model to the data, was set aside, to be treated as a separate topic apart from the foundations of inference. But once the data themselves were considered random variables, the frequentist-Bayesian debate moved into the theoretical world: it became a debate about the best way to reason from random variables to inferences about parameters. This was consistent with developments elsewhere. In other parts of science, the distinction between quantities to be measured and their theoretical counterparts within a mathematical theory can be relegated to a different subject—to a theory of errors. In statistics, we do not have that luxury, and it seems to me important, from a pragmatic viewpoint, to bring to center stage the identification of models with data. The purpose of doing so is that it provides different interpretations of both frequentist and Bayesian inference, interpretations which, I believe, are closer to the attitude of modern statistical practitioners.

A familiar practical situation where these issues arise is binary regression. A classic example comes from a psychophysical experiment conducted by Hecht *et al.* (1942), who investigated the sensitivity of the human visual system by constructing an apparatus that would emit flashes of light at very low intensity in a darkened room. Those authors presented light of varying intensities repeatedly to several subjects and determined, for each intensity, the proportion of times each subject would respond that he or she had seen a flash of light. For each subject the resulting data are repeated binary observations ("yes" perceived vs. "no" did not perceive) at each of many intensities and, these days, the standard statistical tool to analyze such data is logistic regression. We might, for instance, use maximum likelihood to find a 95% confidence interval for the intensity of light at which the subject would report perception with probability $p = .5$. Because the data reported by Hecht *et al.* involved fairly large samples, we would obtain essentially the same answer if instead we applied Bayesian methods to get an interval having 95% posterior probability. But how should such an interval be interpreted?

3 INTERPRETATIONS

Statistical pragmatism involves mildly altered interpretations of frequentist and Bayesian inference. For definiteness I will discuss the paradigm case of confidence and posterior intervals for a normal mean based on a sample of size n, with the standard deviation being known. Suppose that we have $n = 49$ observations that have a sample mean equal to 10.2.

> *Frequentist assumptions:* Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables from a normal distribution with mean μ and standard deviation $\sigma = 1$. In other words, suppose $X_1, X_2, \ldots, X_n$ form a random sample from a $N(\mu, 1)$ distribution.

Noting that $\bar{x} = 10.2$ and $\sqrt{49} = 7$ we define the inferential interval

$$I = (10.2 - \frac{2}{7}, 10.2 + \frac{2}{7}).$$

The interval I may be regarded as a 95% confidence interval. I now contrast the standard frequentist interpretation with the pragmatic interepretation.

> *Frequentist interpretation of confidence interval*: Under the assumptions above, if we were to draw infinitely many random samples from a $N(\mu, 1)$ distribution, 95% of the corresponding confidence intervals $(\bar{X} - \frac{2}{7}, \bar{X} + \frac{2}{7})$ would cover μ.
>
> *Pragmatic interpretation of confidence interval*: If we were to draw a random sample according to the assumptions above, the resulting confidence interval $(\bar{X} - \frac{2}{7}, \bar{X} + \frac{2}{7})$ would have probability .95 of covering μ. Because the random sample lives in the theoretical world, this is a theoretical statement. Nonetheless, substituting
>
> $$\bar{X} = \bar{x} \tag{1}$$
>
> together with
>
> $$\bar{x} = 10.2 \tag{2}$$
>
> we obtain the interval I, and are able to draw useful conclusions as long as our theoretical world is aligned well with the real world that produced the data.

The main point here is that we do not need a long-run interpretation of probability, but we do have to be reminded that the unique-event probability of .95 remains a theoretical statement because it applies to random variables rather than data. Let us turn to the Bayesian case.

> *Bayesian assumptions:* Suppose $X_1, X_2, \ldots, X_n$ form a random sample from a $N(\mu, 1)$ distribution and the prior distribution of μ is $N(\mu_0, \tau^2)$, with $\tau^2 >> \frac{1}{49}$ and $49\tau^2 >> |\mu_0|$.

The posterior distribution of μ is normal, the posterior mean becomes

$$\bar{\mu} = \frac{\tau^2}{\frac{1}{49} + \tau^2} 10.2 + \frac{\frac{1}{49}}{\frac{1}{49} + \tau^2} \mu_0$$

and the posterior variance is

$$v = \left(49 + \frac{1}{\tau^2}\right)^{-1}$$

but because $\tau^2 >> \frac{1}{49}$ and $49\tau^2 >> |\mu_0|$ we have

$$\bar{\mu} \approx 10.2$$

and

$$v \approx \frac{1}{49}.$$

Therefore, the inferential interval I defined above has posterior probability .95.

> *Bayesian interpretation of posterior interval:* Under the assumptions above, the probability that μ is in the interval I is .95.
>
> *Pragmatic interpretation of posterior interval:* If the data were a random sample for which (2) holds, i.e., $\bar{x} = 10.2$, and if the assumptions above were to hold, then the probability that μ is in the interval I would be .95. This refers to a hypothetical value $\bar{x}$ of the random variable $\bar{X}$, and because $\bar{X}$ lives in the theoretical world the statement remains theoretical. Nonetheless, we are able to draw useful conclusions from the data as long as our theoretical world is aligned well with the real world that produced the data.

Here, although the Bayesian approach escapes the indirectness of confidence within the theoretical world, it can not escape it in the world of data analysis because there remains the additional layer of identifying data with random variables. According to the pragmatic interpretation, the posterior is not, literally, a statement about the way the observed data relate to the unknown parameter μ because those objects live in different worlds. The language of Bayesian inference, like the language of frequentist inference, takes a convenient shortcut by blurring the distinction between data and random variables.

The commonality between frequentist and Bayesian inferences is the use of theoretical assumptions, together with a *subjunctive* statement. In both approaches a statistical model is introduced—in the Bayesian case the prior distributions become part of what I am here calling the model—and we may say that the inference is based on what *would* happen if the data *were* to be random variables distributed according to the statistical model. This modeling assumption would be reasonable if the model *were* to describe accurately the variation in the data.

4 IMPLICATIONS FOR TEACHING

It is important for students in introductory statistics courses to see the subject as a coherent, principled whole. Instructors, and textbook authors, may try to help by providing some notion of a "big picture." Often this is done literally, with an illustration such as Figure 2 (e.g., Lovett, Meyer, and Thille, 2008). This kind of illustration can be extremely useful if referenced repeatedly throughout a course.

Figure 2 represents a standard story about statistical inference. Fisher introduced the idea of a random sample drawn from a hypothetical infinite population, and Neyman and Pearson's work encouraged subsequent mathematical statisticians to drop the word "hypothetical" and instead describe statistical inference as analagous to simple random sampling from a finite population. This is the concept that Figure 2 tries to get across. My complaint is that it is not a good general description of statistical inference, and my claim is that Figure 1 is more accurate. For instance, in the psychophysical example of Hecht *et al.* discussed in Section 2, there is no population of

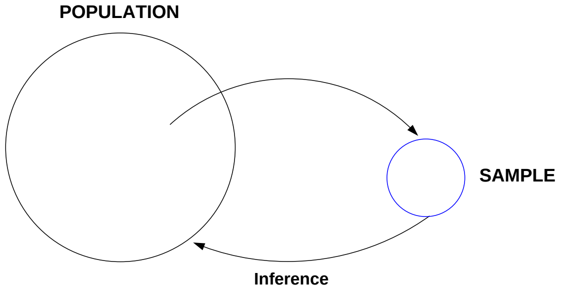

Figure 2: *The big picture according to the standard conception. Here, a random sample is pictured as a sample from a finite population.*

"yes" or "no" replies from which a random sample is drawn. We do not need to struggle to make an analogy with a simple random sample. Furthermore, any thoughts along these lines may draw attention away from the most important theoretical assumptions, such as independence among the responses. Figure 1 is supposed to remind students to look for the important assumptions, and ask whether they describe the variation in the data reasonably well.

One of the reasons the population and sample picture in Figure 2 is so attractive pedagogically is that it reinforces the fundamental distinction between parameters and statistics through the terms *population mean* and *sample mean.* To my way of thinking, this terminology, inherited from Fisher, is unfortunate. Instead of "population mean" I would much prefer *theoretical mean*, because it captures better the notion that a theoretical distribution is being introduced, a notion that is reinforced by Figure 1.

I have found Figure 1 helpful in teaching basic statistics, and I have used it in conjunction with a discussion of statistical thinking, as articulated in Brown and Kass (2009). For somewhat more advanced audiences it is possible to elaborate, describing in more detail the process trained statisticians follow when reasoning from data. A big picture of the overall process is given in Figure 3 (which, in particular, includes the component exploratory data analysis—EDA), but the fundamental message remains unchanged.

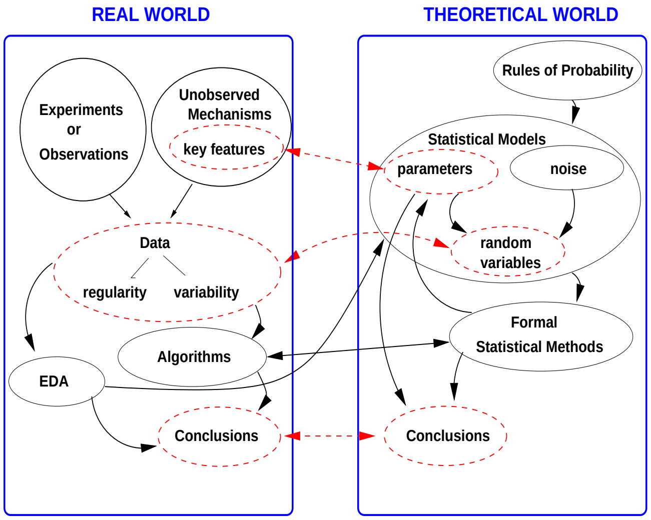

Figure 3: *A more elaborate big picture. Dashed arrows (in red) indicate places where connections between the real and theoretical worlds can not be fully verified.*

5 DISCUSSION

According to my understanding, laid out above, statistical pragmatism has two main features: it is eclectic and it emphasizes the assumptions that connect statistical models with observed data. The pragmatic view acknowledges that both sides of the frequentist-Bayesian debate made important points. Bayesians scoffed at the artificiality in using sampling from a finite population to motivate all of inference, and in using long-run behavior to define characteristics of procedures. Within the theoretical world, posterior probabilities are more direct, and therefore seemed to offer much stronger inferences. Frequentists bristled, pointing to the subjectivity of prior distributions, to which Bayesians responded by treating subjectivity as a virtue on the grounds that all inferences are subjective. While there is a kernel of truth in this observation—we are all human beings, making our own judgments—subjectivism was never satisfying as a logical framework: an important purpose of the scientific enterprise is to go beyond personal decision-making. In fact, the dance around prior distributions has been a bit of a distraction and, it seems to me, the really troubling point for frequentists has been the Bayesian claim to a philosophical high ground, where compelling inferences could be delivered at negligible logical cost. Frequentists have always felt that no such thing should be possible. I believe this feeling has its origins in the gap between models and data, which is neither frequentist nor Bayesian. Statistical pragmatism avoids this irritation by acknowledging explicitly the tenuous connection between the real and theoretical worlds. As a result, its inferences are necessarily subjunctive. We speak of what *would* be inferred if our assumptions *were* to hold. The inferential bridge is traversed, by both frequentist and Bayesian methods, when we act *as if* the data *were* generated by random variables. In the normal mean example discussed in Section 4, the key step involves the conjunction of the two equations (1) and (2). Strictly speaking, according to statistical pragmatism, Equation (1) lives in the theoretical world while Equation (2) lives in the real world; the bridge is built by allowing $\bar{x}$ to refer to *both* the theoretical value of the random variable and the observed data value.

In pondering the nature of statistical inference I am, like others, guided partly by past and present sages (for an overview see Barnett, 1999), but also by my own experience and by watching many colleagues in action. Frequentist and Bayesian logics are different, and some of their methods, especially their hypothesis testing methods, can not be reconciled in any direct way, yet contemporary practitioners move freely between frequentist and Bayesian techniques using one or the other depending on the problem. For example, in a review of statistical methods in neurophysiology my colleagues and I discussed both frequentist and Bayesian methods but did not emphasize this dichotomy (Kass, Ventura, and Brown, 2005). In fact, we never mentioned the distinction between the approaches or their inferential interpretations. Sometimes alternative approaches are used together. In examining several neural firing-rate intensity functions $\lambda^1(t), \ldots, \lambda^p(t)$, assumed to be smooth functions of time t, Behseta *et al.* (2007) developed a frequentist approach to testing the hypothesis $H_0 : \lambda^1(t) = \cdots = \lambda^p(t)$, for all t, that incorporated a Bayesian smoothing method (DiMatteo, Genovese, and Kass, 2001) and its resulting posterior distribution. Such hybrids are not uncommon. This does not mean that a practitioner must walk around with mutually inconsistent interpretations of statistical inference. Figure 1 provides a general framework that encompasses both of the major approaches to methodology while emphasizing the inherent gap between data and modeling assumptions, a gap that is bridged through subjunctive statements. The advantage of the pragmatic framework is that it considers frequentist and Bayesian inference to be equally respectable and allows us to have a consistent interpretation, without feeling as if we must have split personalities in order to be competent statisticians. More to the point, this framework seems to me to resemble more closely what we do in practice: statisticians offer inferences couched in a cautionary attitude. Perhaps we might even say that most practitioners are subjunctivists.

I have emphasized subjunctive statements partly because, on the frequentist side, they eliminate any need for long-run interpretation. For Bayesian methods they eliminate reliance on subjectivism. The Bayesian point of view was articulated admirably by Jeffreys (see Robert, Chopin, and Rousseau, 2009, and accompanying discussion) but it became clear, especially from the arguments of Savage and subsequent investigations in the 1970s, that the only solid foundation for Bayesianism is subjective (see Kass and Wasserman, 1995, and Kass, 2008). Statistical pragmatism pulls us out of that solopsistic quagmire. On the other hand, I do not mean to imply that it really doesn't matter what approach is taken in a particular instance. The Hecht *et al.* example discussed in Section 2 is old and, relative to today's cutting-edge technology, not very interesting. Current attention frequently focuses on challenging, high-dimensional data sets where, in contrast to two-parameter binary regression, frequentist and Bayesian methods may differ. Statistical pragmatism is agnostic on this, and judges by theoretical performance in the context of specific applications.

I called the story about statistical inference told by Figure 2 "standard" because it is imbedded in many introductory texts, such as the path-breaking book by Freedman, Pisani, and Purves (1978) and the excellent and very popular book by Moore and McCabe (1989). My criticism is that the standard story misrepresents the way statistical inference is commonly understood by trained statisticians, portraying it as analogous to simple random sampling from a finite population. As I noted, the population vs. sampling terminology comes from Fisher, but I believe the conception in Figure 1 is closer to Fisher's conception of the relationship between theory and data. Fisher spoke pointedly of a *hypothetical* infinite population, but in the standard story of Figure 2 the "hypothetical" part of this notion—which is crucial to the concept—gets dropped (confer also Lenhard, 2006). I understand Fisher's "hypothetical" to connote what I have here called "theoretical." Fisher did not anticipate the co-option of his framework and was, in large part for this reason, horrified by subsequent developments by Neyman and Pearson. The terminology "theoretical" avoids this confusion and thus may offer a clearer representation of Fisher's idea.[1]

We now recognize Neyman and Pearson to have made permanent, important contributions to statistical inference through their introduction of hypothesis testing and confidence. From today's vantage point, however, their behavioral interpretation seems quaint, especially when represented by their famous dictum, "We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis." Nonetheless, that interpretation seems to have inspired the attitude behind Figure 2. In the extreme, one may be led to insist that statistical inferences are valid only when some chance mechanism has generated the data. The problem with the chance-mechanism conception is that it applies to a rather small part of the real world, where there is either actual random sampling or situations described by statistical or quantum physics. I believe the chance-mechanism conception errs in declaring that data are assumed to be random variables, rather than allowing the gap of Figure 1 to be bridged by statements such as (2). In saying this I am trying to listen carefully to the voice in my head that comes from the late David Freeman (see Freedman and Ziesel, 1988). I imagine he might call crossing this bridge, in the absence of an explicit chance mechanism, a leap of faith. In a strict sense I am inclined to agree. It seems to me, however, that it is precisely this leap of faith that

[1] Fisher also introduced populations partly because he used long-run frequency as a foundation for probability, which statistical pragmatism considers unnecessary.

makes statistical reasoning possible in the vast majority of applications.

Statistical models that go beyond chance mechanisms have been central to statistical inference since Fisher and Jeffreys, and their role in reasoning has been considered by many authors (e.g., Cox, 1990, Lehmann, 1990). An outstanding issue is the extent to which statistical models are like the theoretical models used throughout science (see Stanford, 2006). I would argue, on the one hand, that they are similar: the most fundamental belief of any scientist is that the theoretical and real worlds are aligned. On the other hand, as observed in Section 2, statistics is unique in having to face the gap between theoretical and real worlds every time a model is applied and, it seems to me, this is a big part of what we offer our scientific collaborators. Statistical pragmatism recognizes that all forms of statisical inference make assumptions, assumptions which can only be tested very crudely (with such things as goodness-of-fit methods) and can almost never be verified. This is not only at the heart of statistical inference, it is also the great wisdom of our field.

REFERENCES

Barnett, V. (1999) *Comparative Statistical Inference*, Third ed., Wiley.

Behseta, S. Kass, R.E., Moorman, D., and Olson, C.R. (2007) Testing equality of several functions: Analysis of single-unit firing rate curves across multiple experimental conditions, *Statist. Med.*, 26: 3958–3975.

Brown, E.N. and Kass, R.E. (2009) What is statistics? (With discussion) *Amer. Statist.*, 63: 105–110.

Cox, D.R. (1990) Role of models in statistical analysis, *Statist. Sci.*, 5: 169–174.

DiMatteo, I., Genovese, C.R., and Kass, R.E. (2001) Bayesian curve-fitting with free-knot splines, *Biometrika*, 88: 1055–1071.

Freedman, D., Pisani, R. and Purves, R. (1978) *Statistics* (Fourth edition, 2007), W.W. Norton.

Hecht, Schlaer, and Pirenne (1942) *J. Gen. Physiol.*, 25: 819–840.

Kass, R.E. (2006) Kinds of Bayesians (Comment on articles by Berger and by Goldstein), *Bayesian Anal.*, 1: 437–440.

Kass, R.E., Ventura, V., and Brown, E.N. (2005) Statistical issues in the analysis of neuronal data, *J. Neurophysiol.,* 94: 8-25.

Kass, R.E. and Wasserman, L.A. (1996) The selection of prior distributions by formal rules, *J. Amer. Statist. Assoc.,* 91: 1343-1370.

Lehmann, E.L. (1990) Model specification: The views of Fisher and Neyman, and later developments, *Statist. Sci.*, 5: 160–168.

Lenhard, J. (2006) Models and statistical inference: The controversy between Fisher and Neyman-Pearson, *Brit. J. Philo. Sci.*, 57: 69–91.

Lovett, M., Meyer, O., and Thille, C. (2008) The open learning initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *J. Interactive Media in Education*, 14.

Moore, D.S. and McCabe, G. (1989) *Introduction to the Practice of Statistics* (Fifth edition, 2005), W.H. Freeman.

Robert, C.P., Chopin, N., and Rousseau, J. (2010) Harold Jeffreys' Theory of Probability, revisited (with discussion), *Statistical Science*, to appear.

Stanford, P.K. (2006) *Exceeding Our Grasp*, Oxford.