Submitted to *Journal of Scientific Exploration*

RESPONSE TO DOBYNS

William H. Jefferys

Department of Astronomy
University of Texas at Austin

Dobyns' article suggests some reasons why orthodox statistics might be superior to Bayesian statistics when discussing random event generator statistics. Several of his main arguments are examined and discussed.

Introduction.

I became interested in this topic when, after joining the Society for Scientific Exploration, I ordered the back issues of the *Journal for Scientific Exploration* and set about reading them. While studying the paper of Jahn *et. al.* (1987) I noticed that it actually provided a nice real-life example of the Jeffreys-Lindley paradox. It also made me ask myself why, if the P-values from this research are so small, I had not been moved to regard the psi hypothesis with more favor than I in fact did. This in turn led me to consider more deeply questions of epistemology as viewed through a Bayesian microscope. Some of the issues raised by Dobyns have led me to reexamine these questions, and I believe that the following comments on Dobyns' paper may help others to see these issues more clearly.

I offer the following comments in the spirit of constructive criticism, not negativism. Of course, I cannot conceal my prior, nor do I wish to. I was and remain skeptical of the reality of the paranormal. However, I think that I am typical of scientists who, while quite skeptical, would be willing to change their minds if presented with compelling evidence. Some notion of the kind of evidence that would be compelling, and of where I feel current efforts fall short, are given below.

Choosing an appropriate prior.

Dobyns first investigates a family of priors that are uniform in an interval of width w centered on $p_\Delta = 0.5$. He is following an idea of Lindley (1965, Section 5.6) and shows that such priors approximately replicate the orthodox analysis in the sense that (for any w that is sufficiently large to encompass most of the likelihood function), the Bayesian $100(1-\alpha)\%$ credible interval will include the null only when α is about as small as the classical P-value. But is a uniform prior appropriate to this problem? I contend that it is not. Lindley's idea assumes that we have no particular reason to favor one value of p over another, which is surely not the case here. The PEAR equipment and protocol have been designed to produce an exact 50% hit probability. I therefore have a substantial prior belief in the null hypothesis, whereas Dobyns' prior actually expresses a high degree of skepticism about the null.

According to Dobyns, in the PEAR experiments the maximum artifactual deviation from $p_\Delta = 0.5$ is 1.1×10^{-6}. In using a uniform prior, with $w = 10^{-3}$ (as suggested by

Key words and phrases. Bayesian statistics; Random event generators.

Typeset by $\mathcal{AMS}$-TEX

Dobyns), one would be claiming to be quite certain *a priori* that the null is *false* (with prior odds of about 1000 : 1 *against* the null in this case). This is hardly appropriate, if one has a significant prior belief that the null might be true! Thus, while it may be possible to construct a Bayesian analysis that gives similar results to the orthodox one, in this case it is quite artificial because the prior does not agree with the actual prior of anyone except a person who is already nearly certain that the null is false.

Only a prior that places a substantial proportion of its mass near the null $p_\Delta = 0.5$ can adequately represent the views of a person who has not already made up his mind against the null; in the present case it is quite adequate to approximate this component of the prior as a δ-function. Thus, a prior of the form $\pi_0(p) = a\delta(p - 0.5) + (1 - a)f(p)$, where $f(p)$ is a function representing the prior on the alternative hypothesis, is the only kind that can give due weight to a believable null hypothesis.

This is, of course, the answer to Shafer's objection, that a diffuse prior is being treated as evidence against the hypothesis in question. This is wrong. The diffuse prior expresses skepticism, not about the *hypothesis* in question, but that any *particular* value of the parameter p is the true value required by this hypothesis. But this is exactly what the hypothesis $p \neq 0.5$ says! It, too, is skeptical about any particular value of p, instead regarding p to be a "fudge factor" to be estimated from the data. This is at the heart of the Jeffreys-Lindley paradox. *If one has substantial reasons to believe in a particular value of a parameter as against other values, a parameter-fitting prior that considers all values to be about equally likely is inappropriate.* The case considered here is no different from many similar cases in science. For example, the theory of general relativity predicts a very precise value, $43''$/century, for the perihelion advance of Mercury. Alternative theories that were advanced by nineteenth century astronomers to explain the perihelion advance all contained a "fudge factor" that allowed them to fit virtually any observed perihelion advance. When an observed value turns out to be near the value precisely predicted by a theory (as it did in this case), that theory automatically acquires an *extra measure of credibility* relative to a theory that fits the observed value by resorting to a "fudge factor." Put another way, we want to fit the model to the data without overfitting it. When fitting models, each additional parameter exacts a penalty that must be more than compensated by the increased ability of the model to match the data. Bayesian probability theory allows us to estimate the how big the penalty for adding an additional parameter is (Jeffreys 1939, Bretthorst 1988, Gull 1988). Every scientist agrees with the principle that the number of arbitrary parameters should be kept to a minimum, and that a theory that has fewer parameters is *ipso facto* more credible than a theory with more parameters, *even when the theory with fewer parameters does not fit the data perfectly.*

It has been known for some time that such considerations lead to a Bayesian justification of Ockham's razor. See Jaynes (1979), Smith and Spiegelhalter (1980), Gull (1988), Loredo (1990), Berger and Jefferys (1991), Jefferys and Berger (1991), and MacKay (1991) for discussions. In the PEAR experiments, the unknown value p plays the role of a "fudge factor" that can be adjusted to fit any data compatible with the prior. As a consequence, the hypothesis that some unknown, nonstatistical effect is causing the value of p to differ from 0.5 is *more complex* than the null hypothesis that proposes that $p = 0.5$ to within a very small error. Ockham's razor tells us to favor the simpler theory; the Bayesian calculation tells us just how much the evidence must disagree with the simpler theory before it forces us to favor the more complex one. In this case, the answer is that even a discrepancy of 3.614 standard deviations may not be large enough to force us to favor the

more complex theory, when the effect size $\theta = |p - p_\Delta|/p_\Delta$ is very small.

Whether the discrepancy is large enough to force us to reconsider the simpler hypothesis depends on the width w one chooses for the prior on the alternative hypothesis. Dobyns notes that there is a range of approximately 1000 in the Bayes factors against the alternative. The different values of w correspond to different degrees of specificity in the prior. When w is large, the alternative hypothesis does not make a very specific prediction, and is able to accomodate a wide variety of effect sizes θ without undue pain. Such hypotheses are difficult to falsify on arbitrary data, and are also the least credible after the data are taken. When w is small, the predictions made by the alternative hypothesis are specific, more easily falsified, and therefore more credible. This is seen by the fact that the Bayes factor against the alternative is larger when w is large than when it is small. Thus, Bayes' theorem automatically takes into account the relative complexity or simplicity of the hypotheses (when measured in this way), balancing these against how well each hypothesis agrees with the data. The rub is that one has to choose ones prior on the alternative before looking at the new data. No cheating is allowed!

What are the consequences of assuming a prior that fairly represents real prior belief in the null? Let us consider an extreme case that Dobyns also discusses. This case treats both hypotheses symmetrically, by letting $f(p) = \delta(p - p_0)$, where $p_0 = s/n$ is the value of p that maximizes the likelihood function, and setting $a = 0.5$. Obviously, such a prior is ridiculously favorable to the alternative hypothesis, since it is a maxim of Bayesian and orthodox analysis alike that you should not choose your hypothesis to match the data you have already collected. That would be like being allowed to place your bet after a horse race was run. So this procedure gives us an absolute lower bound on the Bayes factor. Dobyns does the calculation: the result is $B_{min} = 0.00146$. As Dobyns notes, this is already ten times larger than the (one-sided) P-value. I regard this as excellent evidence that the P-value substantially overstates the significance of the PEAR result.

Statistical power.

Dobyns seeks to avoid this conclusion by bringing up the subject of statistical power. Now there are a number of things that can be said about this. First, of course, statistical power is itself an orthodox notion, and is not of much interest in itself to Bayesian analysis. For one thing, it depends upon imagining an ensemble of identical experiments *that have not been run* and considering the frequentist consequences of such experiments. Bayesians regard such an ensemble mythical, and regard speculation based on data sets other than the one actually observed to be vain. But there are other reasons for the Bayesian attitude towards this issue that are not so philosophical.

The first is practical. *Contra* Dobyns, a major reason that Bayesians regard classical P-values as misleading is that real-life experience shows that they are far more likely to reject a point-null hypothesis that happens to be true than their small size would indicate (Lee 1989, pp. 137-38), and that this tendency increases as n gets larger. It is for this reason that Good and others have suggested adjusting P-values, if they must be used, by various correction factors. It is clear that some adjustment, which deflates P-values for large n, is required.

This point concerns the nature of P-values themselves. People used to orthodox thinking are generally unaware that *data-dependent* P-values don't even have a valid frequentist interpretation. To quote Berger and Delampady (1987):

"A Neyman-Pearson error probability, α, has the actual frequentist interpretation

> that a long series of α level tests will reject no more than $100\alpha\%$ of true H_0, but the data-dependent P-values have no such interpretation. P-values do not even fit easily into any of the conditional frequentist paradigms."

Berger and Delampady (1987) give an example to illustrate this point. I paraphrase their argument, which goes as follows:

Suppose that an astronomer hears that many users of statistics rejected null hypotheses at the 5% level when $z = 1.96$ was observed. This astronomer has a typical file drawer full of old experiments involving approximate point nulls, for which the truth eventually became known. Suppose that overall, about half the point nulls turned out to be true, and half false. Our astronomer decides to examine all the cases where the null was originally rejected at or near the exact 5% level, say from $z = 1.96$ to $z = 2.0$. In this subset of tests, where the null was just rejected at the 5% level, the astronomer would discover that the null H_0 would actually have turned out to be true about 30% of the time, which is a far cry from the 5% rejection level.

Berger and Delampady then state the frequentist argument, that if we confine our attention to the sequence of true H_0, then in only 5% of all experiments would $|z| \geq 1.96$. This is true, they agree, but is not the answer we need. What we need to know is what to think about the truth of H_0 when we actually observe a *particular* value of z. In the case of the PEAR data, the particular value $z = 3.614$ has been observed. The P-value is 0.0003, two sided, but as the Berger and Delampady gedanken experiment shows, among a collection of typical experiments that resulted in P-values near 0.0003, the proportion for which the null would actually have been true can be expected to be substantially larger than 0.0003. Again, one is misled by naively looking at P-values.

What this means, of course, is that one cannot "up the ante" after the data are in by choosing the exact P-value as the new rejection level. The proper classical procedure is to choose the rejection level before looking at the data, and then to report either "acceptance" or "rejection" at the predetermined significance level. One must be very careful when interpreting data-dependent P-values.

This is, of course, closely related to a point that Harold Jeffreys made most forcefully when he complained that when one employs tests based on tail-areas, one is rejecting the null hypothesis not only because we happened to have observed an extreme value, but also because we have *not* observed values that are even more extreme. By counting for the null only that part of the tail area that is beyond the observed data, where the curve rapidly approaches zero as $exp(-z^2/2)$, the tail-area test systematically underestimates the actual amount of evidence for the null. I have not seen a satisfactory frequentist answer to his comment (Jeffreys 1939, Section 7.2):

> "If P is small, that means that there have been unexpectedly large departures from prediction. But why should these be stated in terms of P? The latter gives the probability of departures, measured in a particular way, equal to *or greater than* the observed set, and the contribution from the actual value is nearly always negligible. *What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure."

Jeffreys' entire discussion of this point deserves careful reading. Fisher, late in life, came to appreciate the force of Jeffreys' argument. He wrote (Fisher 1956, p. 66),

> "Objection has sometimes been made that the method of calculating Confidence

> Limits by setting an assigned value such as 1% on the frequency of observing 3 or less (or at the other end of observing 3 or more) is unrealistic in treating the values less than 3, which have not been observed, in exactly the same manner as the value 3, which is the one that has been observed. *This feature is indeed not very defensible save as an approximation*" (emphasis added).

Fisher advocated using P-values to suggest interesting areas for future investigation, but using the likelihood function for final analysis. Although he was opposed to Bayesian ideas, it is interesting that he regarded the likelihood function as the firmest foundation for statistical inference. From here it would be but a short step to adopting a fully Bayesian position (although Fisher did not take this step).

Another important point to bear in mind is that when designing a test on frequentist principles, one should choose the value of the parameter for which good power is desired prior to seeing the data. Dobyns' example, which compares the power of the orthodox and Bayesian analyses, is based on his knowledge of the value of the parameter that the data actually indicate ($p = 0.5002$). Dobyns intended this as a pedagogical example; it should not be considered to be a full power analysis, for is not considered good practice to choose the parameter value for the power analysis to be only that where the data ended up. A full power analysis that considered a range of values of p consistent with the results of other investigations would have shown that the Bayesian test in general has extremely good power, although never as good as the classical test.

The final point is that the actual PEAR experiments have been conducted in a quasi-sequential mode. That is, data are gathered for a while, then published. More data are gathered and added to the growing data set. An new analysis is published, and so on. It is guaranteed that if an experiment like this is carried on in a sequential fashion long enough, then with probability 1 there will be occasions when the null hypothesis, even if it is true, will be rejected by a classical test at any arbitrary P-value, no matter how small. Dobyns says that the probability that a terminal Z score of 3.614 could have been attained at any time in the program's history is less than 0.007 (or 0.002 with a different set of assumptions); this is somewhat comforting, but I still have misgivings, since it is based on the same kind of tail-area calculation to which I object.

By contrast, a Bayesian would adopt a rejection criterion (reject the null if at some point its *posterior probability* becomes less than p_0, and reject the alternative if its posterior probability becomes less than p_1) (Berger 1985). Suppose that, unknown to him, the null is actually true. The Bayesian will then have only a small probability of ever rejecting the null, no matter how long he takes data. As a matter of general philosophy, I think it prudent to rely on the safer, simpler Bayesian procedure, even at the price of giving up some statistical power. If this requires the taking of a moderate amount of additional data, so be it.

The need for priors.

The last major complaint that Dobyns lists is the fact that under different priors on the alternative, the Bayesian analysis of these data gives a wide range of Bayes factors. This is certainly true. Dobyns considers it an advantage that the classical analysis gives only one answer.

Again, there are several responses to this. The first response is that the two-sided P-value of 0.0003 from the classical analysis is misleading, as has been pointed out above. It is not the probability that similar true null hypotheses would be rejected, for the reasons that

Berger and Delampady (1987) give. Even with a ridiculously unrealistic prior that gives every advantage to the alternative hypothesis, we have seen that one obtains a Bayes factor that is at least ten times the P-value. Moreover, the Bayes factor has a straightforward interpretation in terms of the probability of the two hypotheses, unlike the classical analysis that tells us something that is irrelevant to the question we are asking. That the fact that classical analysis gives only one answer is no advantage if that answer is misleading or wrong.

Second, prior belief is important, even to orthodox statisticians. To illustrate this point, consider the following variation on Fisher's famous example of the tea-drinking lady (Fisher 1966, pp. 11-25). We consider three hypothetical experiments:

> (1) You have a pack of alphabet cards, one card for each letter. You shuffle them thoroughly and pick a card at random. You show it to a 6-year-old child, who correctly names the letter. You repeat this twice more for a total of three letters in all, and each time the child answers correctly. You ask the child how she is able to accomplish this. She answers that she has learned the whole alphabet watching "Sesame Street."
>
> (2) You take the same pack of cards, shuffle them and look at the card you pick, but do not show it to a subject. The subject correctly names the card. You repeat the process twice more, and each time the card is correctly named. You ask the subject how he is able to do this. The subject answers that he is a professional magician and is able to give you the illusion that he has read your mind using his conjuring skills.
>
> (3) With the same situation as in case (2), the subject answers that he is a psychic and able to read your mind.

What is the difference between these situations? The statistical evidence is the same, but I think that nearly everyone would assess the likelihood of the three explanations differently. Most would be convinced that the child is probably speaking the truth, and many would likewise believe that the magician had the skills he claimed, whereas I believe that most people would demand much more evidence before they would be convinced by the self-proclaimed psychic. Why this difference? The answer is that our prior probability in the three cases is different. We have much experience that 6-year-old children frequently know the alphabet, and a fair amount of experience that magicians can perform surprising feats by the use of trickery and misdirection; however, most people have little evidence that psychic powers are real, so the prior probability that an individual is a genuine psychic is very small. Thus, from a statistical point of view, the same evidence does not necessarily result in the same posterior belief about the claims made.

Once one admits that prior information is relevant in statistical inference, it seems to me that one is led inevitably to accept Bayesian premises. Classical statistics was invented to make statistical inference "objective." In fact, classical statistics is no more objective than Bayesian statistics, but by hiding its subjectivity it gives the illusion of objectivity. As Box (1980) writes:

> "In the past, the need for probabilities expressing prior belief has often been thought of, not as a necessity for all scientific inference, but rather as a feature peculiar to Bayesian inference. This seems to come from the curious idea that an outright assumption does not count as a prior belief....I believe that it is impossible logically to distinguish between model assumptions and the prior distribution of the parameters."

And, more pithily, Good (1973):

> "The subjectivist states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science."

What, then, are we to make of the fact that the Bayes factor for the PEAR data varies over a range of 1000, depending on the choice of prior on the alternative hypothesis? Does this really point to a defect of in the Bayesian approach? I think not. The immediate reason for this variation of the Bayes factor, of course, is clear: the effect, if any, is extremely small. If the effect were substantially larger, then the the Bayes factor would range over a much smaller interval, and (on the same number of data points) any Bayesian who holds one of the priors I have advocated would be quite emphatic in rejecting the null. Thus, Bayesian analysis tells us that it takes a great deal more evidence to convince us of the reality of a very tiny effect than of a large effect.

In my view this should be taken as a *warning: because the effect is so small, these data may not yet provide convincing evidence that an anomaly exists.* Different Bayesian observers, all with priors that are reasonable given their state of prior knowledge, get different results. This warning is clear to everyone who accepts the Bayesian analysis. It is, however, a warning that the classical analysis fails to sound. For this reason, I view it as evidence of a shortcoming in orthodox statistics, and not evidence of a problem with the Bayesian approach.

Future prospects.

The official position of the PEAR project is that they are studying anomalies, not the paranormal. Anomalies may be due to any cause, whether mundane or paranormal. Yet it is obvious that one of the reasons why the PEAR results have excited so much interest, particularly amongst the public, is the possibility that they may have paranormal causes.

Is it possible for experiments such as these to provide evidence for paranormal effects, as opposed to mundane ones? This is an interesting question. As I see it, there is an essential difficulty, because experiments of the kind discussed here cannot discriminate between the two hypotheses. However interesting the anomalies produced by these experiments may be, they cannot tell us whether the anomalies indicate new physics (for example) or something less exciting such as an undetected mundane error.

The problem is that statistics cannot easily discriminate between hypotheses that have essentially the same likelihood function. Since the hypothesis of paranormal effect and the hypothesis of mundane error are equally good at predicting the existence of an anomaly, statistics cannot tell us which one is right. And since the posterior probability is proportional to the likelihood times the prior, new data cannot much alter our opinion about the relative merits of each hypothesis. No amount of statistical analysis can change this situation: the only cure is to do a different kind of experiment, one that *can* distinguish between these two hypotheses.

Consider, for example, the the card guessing experiment that I presented earlier. I deliberately chose the numbers (3 cards of an alphabetic set) so that, were the subject guessing, the probability of getting all three correct would approximately equal the classical P-value from the PEAR data. What of subject number 3, the one who claims to have psychic powers? Perhaps this subject is actually a conjurer like subject 2, but is pretending to have psychic powers. We know that conjurors can accomplish by mundane means that which appears to be paranormal. We know that magicians sometimes pose as psychics. The

prior probability that a surprising feat like card-guessing was accomplished by mundane means seems to me to be much larger than that it was accomplished by paranormal powers. But the experiment itself does not allow me to distinguish between the two hypotheses, so it is unlikely to alter my opinion that paranormal effects are not involved by much. I do not need to know exactly how the feat was accomplished in order to reach this conclusion. (This point has been made by Jaynes 1990, and discussions by Good 1950, and Mosteller and Wallace 1964 are also relevant.)

The smallness of the signal in the REG experiments is not the problem. The signals in many experiments in physical sciences are far smaller than those claimed to exist in the REG work, but the evidence that these signals are real is often absolutely compelling. For example, for 20 years, the University of Texas' McDonald Observatory has been beaming short pulses of laser light at a reflector on the Moon. About one time in ten, a single photon returns and is detected. It has become routine to detect and identify that single photon from amongst many thousands of "noise" photons that also enter the telescope. This was accomplished by careful experimental design and clever technique. If the experiment depended on the kind of $\sqrt{N}$ "beating down the noise" that has become the norm in parapsychological research, the laser ranging experiment could not work.

I believe that it would be interesting try to devise experiments based on very different principles from the ones that have been conducted since Rhine introduced the statistical/cognitive-science model into parapsychological research. Modern technology has made available many devices that might be pressed into service. For example, a recent report (Eigler *et. al.* 1991) describes a switch that can be turned on and off by moving a single atom across a microscopic gap. If PK can really affect the roll of dice, or the fall of balls in a random mechanical cascade, it might be capable of moving a single atom across a gap of a few microns. If PK is a real phenomenon, the principle behind this switch might make it possible, for example, to build a device that would let me turn my TV set on or off at will just by my thinking about it. If this were possible, it would be very exciting indeed: PK as a phenomenon would become as commonplace and uncontroversial as electricity. Of course, the experiment might well fail. But even in this case, one would be better off for having done the experiment, because it would enable one to rule out certain models of how PK, if real, might work. This, in turn, might suggest other lines of research.

Conclusions.

To turn Richard W. Hamming's phrase, the purpose of statistics is insight, not numbers. Statistics is a tool for helping us to make sensible decisions in the face of data, decisions that are consistent with our prior knowledge and with new information that may come to our attention. It is not a tool for bludgeoning those who disagree with you, either with small P-values or with large Bayes factors. The statistician can provide guidance as to what the statistics mean; but the individual consumer of the statistics remains the ultimate judge of whether the evidence of any experiment is convincing.

Statistics cannot substitute for good judgement, nor can it transform a flawed experiment into a valid one. Where an experiment cannot distinguish between two equally capable explanations, no amount of statistical analysis will change that situation. Where data are at the margins of detectablility, the solution is to design a better experiment, not more statistics. P-values, as provided by orthodox statistical methods, can be and often are misunderstood even by those who use them every day. Data-dependent P-values contain subtle traps that makes their interpretation hazardous. Bayesian statistics, because

of its straightforward interpretation, and because the assumptions are out in the open, offers a way to clarify and sharpen our thinking about experiments, and by giving us new insight about why parapsychological experiments are not having their intended effect of convincing a skeptical scientific world, they can point out research directions that might be more fruitful.

Acknowledgements.

I thank James O. Berger and York Dobyns for comments on earlier drafts of this article. The opinions expressed here, and any errors, are of course my own responsibility.

References

Berger, James O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Second Edition. New York: Springer–Verlag.

Berger, James O. and Delampady, Mohan. (1987) "Testing precise hypotheses." *Statistical Science* **2**, 317-352.

Berger, James O. and Jefferys, William H. (1991). "Minimal Bayesian testing of precise hypotheses, model selection, and Ockham's razor." Technical Report, Purdue University.

Box, G.E.P. (1980). *J. Roy. Statist. Soc. (Ser. A)* **143**, 383-430. (Quoted in Berger 1985).

Bretthorst, G. Larry (1988). *Bayesian Spectrum Analysis and Parameter Estimation. Lecture Notes in Statistics Series Vol. 48*. New York: Springer-Verlag.

Eigler, D.M., Lutz, C.P., and Rudge, W.E. (1991). "An atomic switch realized with the scanning tunnelling microscope." *Nature* **352**, 600-603.

Fisher, Ronald A. (1956). *Statistical Methods and Scientific Inference*. New York: Hafner Publishing Company.

Fisher, Ronald A. (1966). *The Design of Experiments*, 8th Edition. Edinburgh: Oliver and Boyd. (Quoted in Lee 1989).

Good, I.J. (1950). *Probability and the Weighing of Evidence*. London: Charles Griffin & Co. pp. 68-71, 81-82.

Good, I.J. (1973). in *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott (eds.) Toronto: Holt, Rinehart & Winston. (Quoted in Berger 1985).

Gull, S. (1988). "Bayesian inductive inference and maximum entropy. In G.J. Erickson and C.R. Smith (eds.), *Maximum Entropy and Bayesian Methods in Science and Engineering (Vol 1), 53-74*. Dordrecht: Kluwer Academic Publishers.

Jahn, R.G., Dunne, B.J., and Nelson, R.D. (1987). "Engineering anomalies research," *Journal of Scientific Exploration* **1**, 21-50.

Jaynes, E.T. (1979). "Inference, method, and decision: Towards a Bayesian philosophy of science." *Journal of the American Statistical Association* **74**, *740-41.*

Jaynes, E.T. (1990). *Probability Theory—The Logic of Science*, in press, Chapter 5.

Jefferys, W.H. (1990). "Bayesian analysis of random event generator data." *Journal of Scientific Exploration* **4**, 153-169.

Jefferys, W.H. and Berger, James O. (1991) "Sharpening Ockham's razor on a Bayesian strop." To appear in *American Scientist*.

Jeffreys, H. (1939). *Theory of Probability, Third Edition*. Oxford: Clarendon Press.

Lee, Peter M. (1989). *Bayesian Statistics*. New York: Oxford University Press.

Lindley, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint.* Cambridge: Cambridge University Press.

Loredo, T.J. (1990). "From Laplace to Supernova 1987A: Bayesian inference in astrophysics." In P. Fougere (ed.), *Maximum Entropy and Bayesian Methods, 81-142*. Dordrecht: Kluwer Academic Publishers.

MacKay, David J.C. (1991). "Bayesian Interpolation." Submitted to *Neural Computation*.

Mosteller, Frederick and Wallace, David L. (1964). *Applied Bayesian and Classical Inference: The Case of The Federalist Papers.* 2nd Edition. New York: Springer-Verlag. pp. 88-91.

Smith, A.F.M. and Spiegelhalter, D.J. (1980). "Bayes factors and choice criteria for linear models." *J. Royal Statist. Soc. B* **42**, *213-220.*