Journal of Scientific Exploration, Vol. 4, No. 2, pp. 153-169, 1990
0892-3310/90 $3.00+.00
Pergamon Press plc. Printed in the USA.
©1991 Society for Scientific Exploration

Bayesian Analysis of Random Event Generator Data

WILLIAM H. JEFFERYS

Department of Astronomy, University of Texas at Austin. Austin, TX 78712

Abstract—Data from experiments that use random event generators are usually analyzed by classical (frequentist) statistical tests, which summarize the statistical significance of the test statistic as a *p*-value. However, classical statistical tests are frequently inappropriate to these data, and the resulting *p*-values can grossly overestimate the significance of the result. Bayesian analysis shows that a small *p*-value may not provide credible evidence that an anomalous phenomenon exists. An easily applied alternative methodology is described and applied to an example from the literature.

Introduction

In recent years a new type of experiment using a random event generator (REG) has become popular in parapsychological research (Schmidt, 1970; Jahn, Dunne, & Nelson, 1987). This methodology is a modern refinement of the VERITAC technology of Smith, Daglen, Hill, & Mott-Smith (1963), which itself embodies features of Tyrrell's (1936) experiments. The technique is based on an electronic device driven by a source of random events (usually a radioactive source or an electronic device that produces Johnson noise), arranged so that one gets a random sequence of 0s and 1s with theoretically equal probability for each outcome. The outcomes are counted and recorded automatically so as to reduce the possibility of tampering and human error. The statistical properties of the REG are investigated and used to establish a baseline. A subject then attempts to "influence" the REG and obtain a sequence of pulses with a distribution that is different from the baseline. Typically the experimental protocol has control sequences interspersed among the experimental ones.

The results of such experiments appear at first glance to provide very strong evidence against the null hypothesis, that the subject has no effect on the REG. A number of published experiments display small, but highly significant deviations between baseline and experimental runs, and the null hypothesis is rejected with very small *p-value*s. However, despite these apparently strong results, such experiments have by and large not moved skeptics to change their opinions. This is an apparently paradoxical situation that calls for an explanation.

Acknowledgments. The author would like to thank James E. Alcock, James O. Berger, and Ray Hyman for reading earlier drafts of the manuscript and providing valuable comments, and the referees for their careful reviews and detailed suggestions.

Some critics of the REG work have pointed out methodological errors (e.g., Alcock, 1990; Hansel, 1989), whereas others have suggested that the deviation from chance, while highly significant in a statistical sense, is nevertheless far too small to be of practical interest. As Cicchetti (1987) remarks, "We have a highly significant statistical result, which is of utterly trivial consequence from a clinical or substantive point of view." Yet others (e.g., Berger & Sellke, 1987; Navon, 1987; Utts, 1988) have pointed out that in real life a point null hypothesis will always be rejected if enough data are taken, because there will inevitably be uncontrolled sources of bias. On the other hand, Palmer and Rao (1987) have countered that any such bias would be insignificant. Others have objected that a tendency not to publish inconclusive studies would lead to an excess of published studies with small *p*-values. Based on a meta-analysis of published studies, Radin, May, and Thomas (1986), and Radin and Nelson (1989) have responded that the number of unpublished experiments cannot be large enough to explain the published results.

In this paper I will consider a different issue. There is yet another difficulty with REG studies that can result in grossly overestimating their statistical significance, even if we take the data at face value. In many ways the success of the REG approach is its ability to provide a very large body of data based on numerous trials in a relatively short time. Ironically, this also turns out to be a potential Achilles's heel when viewed from a statistical point of view.

The Jeffreys-Lindley Paradox

A result that has been known to statisticians for a very long time bears directly on this question. Known as Jeffreys's paradox (or Lindley's paradox), it has been succinctly described in a beautiful paper by Lindley (1957). (See also Shafer, 1982; Berger & Sellke, 1987; and Berger & Delampady, 1987, and their respondents, for extended discussions of the Jeffreys-Lindley paradox.) In his paper, Lindley constructs an example that shows that if H is a simple hypothesis and x the result of an experiment, the following two phenomena can exist simultaneously:

1. a significance test for H reveals that x is significant at the level $p < \alpha$, where $\alpha > 0$ can be as small as we wish, and

2. the posterior probability of H, given x, is, for quite small prior probabilities of H, as high as $(1 - \alpha)$.

This means that the classical test can reject H with an arbitrarily small p-value, while at the same time the evidence can convince us that H is almost certainly true.

At first sight it might appear that Lindley's example is simply a pathological result of little practical interest, but the result is in fact quite general and can apply to almost any reasonable classical significance test. An important characteristic of this result is that the greater the number of trials, the more likely it is that the conditions of the "paradox" will be fulfilled. This is a consequence of the law of the iterated logarithm (Feller, 1957, pp. 191-197), which implies that given any target value of the test statistic Z, no matter how large, (or equivalently, any target p-value α, no matter how small) the actual value of Z will inevitably exceed the target for certain large numbers of trials, even if the null hypothesis is true. Since the REG paradigm depends on large numbers of trials to attain the high levels of statistical significance that are reported, this is obviously relevant.

Lindley explicitly noted the significance of this result for parapsychology, as have others (e.g., Good, 1982; Berger & Delampady, 1987). However, these warnings do not seem to have received much attention within the parapsychological community. With the exception of Navon (1987) and Utts (1988), I have not seen references to Bayesian hypothesis testing in the parapsychological literature, and have found no references whatsoever to the Jeffreys-Lindley paradox.

Bayesian Hypothesis Testing

The field of Bayesian hypothesis testing has received a good deal of attention in recent years as statisticians have come to recognize difficulties with classical (frequentist) hypothesis testing (Berger & Berry, 1988; Howson & Urbach, 1989). In this paper I will give only a brief introduction to Bayesian analysis, and will quote the necessary results where appropriate. The interested reader is directed to an important paper by Edwards, Lindman, & Savage (1963) for a lucid introduction to the topic from the point of view of using it in practical problems. Intermediate-level introductions can be found in Lee (1989), and Press (1989). More advanced treatments can be found in Zellner (1971), Savage (1972), and Berger (1985).

An important feature of Bayesian hypothesis testing is that the analysis is not affected by considerations such as stopping rules. The intentions of the investigator are simply irrelevant to a Bayesian. A properly formulated Bayesian hypothesis test will be relatively immune even to an obviously illegitimate stopping rule designed to "fool" the analysis, such as the strategy of *sampling to a foregone conclusion*: "Stop when the value of the test statistic exceeds a preassigned number k" (Berger, 1985, §7.7). This is an important consideration, because many criticisms of statistics as they have been used in parapsychology have objected to optional stopping, which has some times been employed, particularly in older studies. It is a fact that whenever the testing protocol is "open-ended," such that more data can be added to the database at will either by adding subjects or by conducting additional trials on a given subject, then classical hypothesis testing is made enormously more complicated (Berger & Berry, 1988). Under these conditions it is plainly wrong to use Z-values to infer p-values in the usual way. The fact that Bayesian analysis is not affected by such considerations is a powerful argument in its favor, since it makes questions about optional stopping moot.

As Good (1982) emphasizes, the physical evidence is the same whether the experimenter was forced to stop an experiment or chose to stop, and so the stopping rule, whatever it is, ought not to have any effect on the correct logical evaluation of evidence. The fact that classical tests can give different results on the same physical evidence, depending upon the intentions of the experimenter, is very troublesome. Since Bayesian analysis is not affected by the stopping rule, or even by changes in the stopping rule based on data that may become available partway through the study, the experimenter who adopts Bayesian methods can with a clear conscience decide to alter the protocol, for example, to take additional data on a promising subject and add them to the corpus of data, without compromising the statistical integrity of the experiment. This positive feature of Bayesian analysis has both practical and ethical implications in other fields (Berger & Berry, 1988). For example, when testing a new drug, it is ethical to withhold the experimental treatment from patients in the control group if, partway through the test, clear evidence of its superiority over the alternative treatment were found? To break the protocol might

invalidate the whole study, if the data are to be analyzed classically; but to withhold the better treatment from seriously ill patients who might otherwise benefit from it might be equally unacceptable. A similar (though not life-threatening) ethical dilemma could face a parapsychologist who discovered a very promising subject partway through a study.

We are interested in testing a null hypothesis H against an alternative hypothesis $H' = \overline{H}$. In our case the null hypothesis is the obvious one, that the results of our experiment are due entirely to chance under some assumed probability law. For an ideal REG, 'chance' means that the probability is $p_0 = 0.5$ that a particular trial results in a 1, and $q_0 = 1\text{-}p_0 = 0.5$ that the result is 0. The alternative hypothesis is somewhat more nebulous. If the subject is actually able to influence the REG, then he has presumably accomplished this by altering the true probability of success for a single trial from $p_0 = 0.5$ to some other value $p \neq p_0$. However, we do not know precisely what this other value is; it is in any case known only vaguely, and (if the subject were extremely effective) it might conceivably lie anywhere in the interval [0,1]. Values less than 0.5 correspond, of course, to "psi missing," and are often considered significant by parapsychologists.

The Bayesian solution to the problem of testing a null hypothesis H against an alternative hypothesis H' is given by Edwards *et. al.* (1963, p. 220). (Here it is assumed that H and H' constitute a mutually exclusive and exhaustive set of outcomes for the experiment.) According to Bayes's theorem, the posterior odds in favor of the null hypothesis, given the observed data x, are

$$\Omega(H \mid x) = \frac{p(x \mid H)}{p(x \mid H')}\Omega(H) \tag{1}$$

where $p(x \mid H)$ is the probability of obtaining x, given that the null hypothesis is true. and $p(x \mid H')$ is the probability of obtaining x, given that the alternative hypothesis is true. $\Omega(H)$ is the prior odds in favor of H, and must be assigned by the user. Equation (1) can be rewritten as:

$$\Omega\,(H \mid x) = B\;\Omega\,(H),$$

where B is called the *likelihood ratio* or *Bayes factor* for the problem. Explicitly, we have

$$B = \frac{p(x \mid H)}{p(x \mid H')}.$$

To use Eq. (1), the user must assign a value for the prior odds $\Omega\,(H)$. This is a number that represents his confidence, prior to knowing x, in the proposition that the null hypothesis is true, as against the proposition that the alternative hypothesis is true. It is easiest to think of this as the odds that a bettor would be willing to give (or take) when betting for (or against) the null hypothesis, against an opponent who takes the other side of the bet (but see Howson & Urbach, 1989, for a detailed discussion). If the user's prior opinion about the two hypotheses were completely neutral, for example, he would pick $\Omega(H) = 1$. On the other hand, if he had some prior confidence in the null hypothesis, he would pick a number $\Omega(H) > 1$ equal to the odds at which he is just willing to bet against the alternative hypothesis. It can be shown (Savage, 1972) that this choice of Ω is optimal, in the sense that at those odds, the bettor would have no particular reason to prefer either

side of the bet, whereas at any other odds he would believe that he would have an advantage by taking one or the other side.

Utts (1988) remarks that "most researchers have strong opinions about the probability that psi is real, and these opinions play a central role in how psi researchers and critics evaluate the evidence." The Bayesian approach allows us to recognize these strongly held opinions explicitly and to take them into account in the analysis. A skeptic who initially believed quite strongly in the null hypothesis would, and in fact should pick as her prior odds $\Omega(H) \gg 1$, whereas someone who was strongly predisposed towards the alternative hypothesis ought to pick $\Omega(H) \ll 1$. Some people would probably choose the former, following the maxim "extraordinary claims require extraordinary evidence" (Abelson, 1978). Others would choose the latter, and still others, being relatively neutral, would pick $\Omega(H) \approx 1$. But never mind. When applied to the problem discussed in this paper, the Bayesian approach guarantees that the correct decision will ultimately prevail, regardless of one's prior opinions. However, much more data may have to be taken to convince some observers than others. And this is, of course, as it should be.

Sometimes it is easier to think in terms of probabilities, rather than odds, although in this paper the odds are simpler to work with since they make it easy for various individuals having very different prior odds to compute their posterior odds. The probability can easily be computed from the odds using the formula.

$$p = \frac{\Omega}{1+\Omega}.$$

We wish to test a point null hypothesis, which is a classical problem in statistics. However, we will now consider this problem from a Bayesian point of view. The factor in the numerator of the Bayes factor is given by the binomial distribution. The denominator, however, is more problematical. As a first approximation consider the situation of a completely naive individual who has no prior opinion at all about how large an effect might be reasonable, that is, one who believes that a large effect (in either direction) is just as likely as a small one. This would be reflected by a prior probability distribution that is uniform on [0,1]. Such a choice is noncommittal in the sense that all values of p are considered equally likely if H' is true. However, most people will probably have a prior opinion that some values of p are more likely than others. For example, values of p that are closer to 0.5 would probably be considered more probable than values that are far from it. No matter, we can formulate the problem in such a way that the user of the statistics (i.e., the reader) is in full control of these hypotheses. The important point is that regardless of the different choices of prior probability made by different observers (provided certain very mild conditions are satisfied), the consistency of Bayes's rules guarantees that ultimately all observers will agree on the result, in the sense that in the limit of arbitrarily many observations the posterior probability will go to zero or to one for all.

In the case of Bernoulli trials, such as are generated by a REG, we have from Edwards, Lindman, and Savage's Eq. (10) the result

$$B = \frac{p_0^r (1-p_0)^{n-r}}{\int_0^1 u^r (1-u)^{n-r} \phi(u \mid H') du} \tag{2}$$

where p_0 is the probability of success for a single trial under the null hypothesis, n is the total number of trials, and r is the number of successes. The function ϕ is the prior probability density of p under the alternative hypothesis H', and must be chosen by the user. In our situation, if prior to taking data we thought that all values of $p \in [0,1]$ were equally likely under the alternative hypothesis, then ϕ would be identically 1. If, on the other hand, we initially thought that values of p close to p_0 were more likely, then the density ϕ would be peaked at p_0 with a width that represented the range of values of p that we thought probable. Again, for subjective reasons, different users will pick different functions ϕ. Thus, the assumptions remain under the full control of the user.

The reader will have noticed that users of the Bayesian approach have to express their subjective opinions about the relative merits of the two hypotheses prior to considering the data. They may also have to make their beliefs about the alternative hypothesis more explicit than simply that it is "not the null hypothesis," which would suffice in the case of a classical test. In other words, Bayesian analysis forces us to be much more explicit about our prejudices than does classical hypothesis testing. This may seem like a weakness of the Bayesian approach, because it involves us in subjective assumptions. In fact, it is a strength, because the classical approach also has assumptions, although they are not often stated. Since classical statistical analysis doesn't explicitly bring these assumptions out into the open, it is easy to be misled into thinking that it is completely objective. Such common usages as "the null hypothesis is rejected ($p < 0.01$)," clearly betray this prejudice. But as Berger and Berry (1988) point out, the classical approach really provides only an "illusion" of objectivity, without its substance. I shall have more to say about this later. For the moment, however, I will simply note that we are particularly interested in the Bayes factor. It tells us how strongly our opinions ought to change after the data are in, as compared to before.

An Example

In this section I will apply a Bayesian analysis to the data of Jahn *et. al.* (1987), generally considered one of the better of the REG experiments (although this work is not above criticism: see Alcock (1990 pp. 104-110). I will not attempt to analyze all of their results, but will instead deal only with the combined figure given in their Table 1 in the row labeled "all." The reasons for this are twofold. First, it is the combined result that has been quoted most frequently. Second, a Bayesian analysis is rather more involved than the corresponding classical analysis, because more things have to be made explicit. In order to take into account the subjective needs of various observers, we will find that a single number (like a p-value) is insufficient to summarize the statistical situation. Once we have seen how to analyze this particular case, however, we can apply the same ideas to other data of interest.

Let us therefore consider the results in the ΔPK column. This results from adding the success score when the intention of the operator was to score high (PK+) to the success score when the intention was to score low (PK-). There were 522,450 "runs" each consisting of 200 Bernoulli trials, giving a total of 104,490,000 trials in all. The Z-score was 3.614 standard deviations relative to the null hypothesis $p_0 = 0.5$. This corresponds to an excess of $3.614 \times (0.25 \times 104{,}490{,}000)^{1/2} = 18471$ successes over chance expectation, or about 0.018%. Nevertheless, even though this excess is very small, the null hypothesis is decisively rejected ($p < 0.00015$: one-tailed; $p < 0.0003$: two tailed) because n is so large. The question is, what would a Bayesian analysis of the same data tell us?

As a first approximation, let us consider the case of a uniform prior under the alternative hypothesis, $\phi(u \mid H') = 1$ (which corresponds to a two-sided alternative). Then $r = 52{,}263{,}471$, and upon performing the calculation of Eq. (2) we find that the Bayes factor is $B = 12$; that is, whatever the prior odds $\Omega(H)$, anyone who makes these assumptions should, after considering these data, give odds against the alternative hypothesis that are 12 times *greater* than before. Thus, this simple Bayesian analysis actually leaves such an observer more confident of the null hypothesis than before*, despite the strong rejection of the null hypothesis by the classical test!* This example demonstrates how the Jeffreys-Lindley paradox may suddenly become very relevant indeed.

We will probably want to choose a nonuniform prior in preference to the uniform prior of this simple example. For example, if we believe that the subject's influence would be likely to be relatively modest, we would pick a prior ϕ that concentrates its mass near 0.5, and which tails off as we move away from that value. A convenient family of priors that reflects these considerations consists of functions of the form $\phi(u \mid H') = C_k[u(1 - u)]^k$, where C_k is a normalizing constant and $0 \le k < \infty$. For $k = 0$, ϕ is the uniform prior. As k gets larger and larger, these functions become more and more peaked and approach normal distributions with mean 0.5 and variance $1/(4(2k + 3))$. Even for small k the normal approximation is a pretty good one, concentrating approximately 2/3 of the mass within one standard deviation and most of the remaining mass within 2 standard deviations. For example, with $k = 4$ ($\sigma = 0.15$), 65.9% of the mass is contained within 1 standard deviation and 96.2% within 2 standard deviations, as compared with 68.3% and 95.5%, respectively, for the normal distribution. The interpretation of the results is, therefore, relatively straightforward. Finally, this family of priors is easy to evaluate and integrate (the integral is a beta function). Note that this family of priors corresponds to a two-sided alternative, and should be compared with the results of a two-sided classical test.

Letting σ be the standard deviation of ϕ, we can evaluate the Bayes factor as a function of $a = 0.5 + \sigma$, where $0.5 < a \le 1$ can be picked at will by the user. When the normal approximation is satisfactory (say, $a \lesssim 0.65$), this prior concentrates about two-thirds of its mass within the range $[1 - a, a]$. In general it is conservative to choose a larger, rather than a smaller value of a, so as not to make the prior distribution so narrow that it would virtually exclude effects that are considered likely. Note also that a should be chosen prior to considering the data.

The results of this analysis for the data of Jahn *et. al.* are shown in the upper curve of Figure 1 . In reading the graph, one should think in terms of standard deviations. Thus, if one wants to choose a value of a that concentrates about 95% of the prior probability (i.e., approximately two standard deviations) within the interval [0.4, 0.6], one would choose $a \approx 0.55$. Also, one who prefers to think in terms of probabilities instead of the Bayes factor can calculate the posterior probability by reading the Bayes factor off the graph and using

$$p(H \mid x) = \frac{1}{1 + (1 - \pi_0)/\pi_0}$$

where π_0 is the prior probability of H. As a practical rule, when B is small, less than 0.05 or so, the posterior probability is approximately equal to B in the neutral case that $\pi_0 = 0.5$.

Notice that for all but a small range of a near 0.5, the Bayes factor is greater than 1, which means that an observer who picks $a \geq 0.534$ would believe that the data provide evidence for, not against, the null hypothesis. Only for a less than 0.534 does the Bayes factor become less than unity, and only in a very tiny interval near $a = 0.5$ is it significantly less than unity. A blowup of the region near $a = 0.5$ is shown in Figure 2. Again consulting the upper curve, we see that the region where the Bayes factor is less than 0.05 (which begins to provide moderate evidence against the null hypothesis) is very narrow indeed, extending only out to about 0.5017 (corresponding to an effect of 0.34%)

If we happened to choose in advance the particular value of a that minimizes the Bayes factor, which is most unlikely since this value is not fixed but is a function of n and Z, we would find that B, and the corresponding posterior probability (on $\pi_0 = 0.5$) are *at least* 0.0088, which is some 30 times larger than the p-value at which the two-sided classical test rejects the null hypothesis. If we follow Berger and Sellke (1987) and choose, from amongst the natural class of priors that are symmetrical about 0.5 and monotone nonincreasing as one goes away from 0.5, the one that is most favorable to the alternative hypothesis, we would obtain a minimum of about 0.0064, which is still some 20 times larger than the classical p-value. Thus, the classical test overestimates the significance of the result by at least a factor of 20, and after considering these data, most Bayesian observers would probably conclude that the real factor is in the hundreds or even the thousands. So we see that even under quite favorable circumstances, the classical test greatly exaggerates the strength of the result. Many would probably conclude that these data provide at best mild evidence against the null hypothesis, and some would feel that the data actually support it.

Jahn *et. al.* actually report their results in terms of a one-tailed test. For reasons that I will discuss in the next section, I believe that a one-tailed test is not appropriate for this kind of experiment. However, in Figures 1 and 2 the lower curve shows the results of applying a one-sided analog of the above test, using the family of prior distributions $\phi' = 2\phi\Pi(2u - 1)$, where $\Pi(x)$ is the "box" function that is 1 on [0,1] and 0 elsewhere, and $\phi = \phi(u \mid H')$ was defined previously. This sets the prior probability on the alternative hypothesis equal to zero for $p < 0.5$, which means that we are disregarding "psi missing." For easy comparison with the two-sided case, the curve for the one-sided case is plotted against the same value of a for the function ϕ that corresponds to the two-sided case. The results are similar to the analysis for the two-sided alternative, except that as we move away from $a = 0.5$, the Bayes factor takes on values that are almost exactly half the previous values. The points where B just exceeds the values 1 and 0.05 are moved to 0.566 and 0.5034, respectively. Since the one-tailed p-value is likewise half of its two tailed value, we see that the classical test still overestimates the strength of the result by about the same factor as before.

In choosing the prior on the alternative hypothesis, one should be guided by any prior knowledge one has. For example, I would personally want to take into account the fact that effects of 5%, 10%, 20%, and even larger have been reported in the parapsychological literature for REG experiments (e.g., Schmidt, 1969; Schmidt & Pantas, 1972). Thus, I would tend to prefer a prior that would allow for a fairly substantial effect, perhaps 10% ($a \approx 0.55$) or more. Others, with different experience and information, may come to different conclusions.

Regardless of the choice of a, we are now in a position to see what the effect of these results would likely be on the attitudes of different individuals. The skeptic, who has a prior odds ratio $\Omega(H) \gg 1$, is probably not going to be convinced that an anomaly exists. Depending on the value of a that he chooses, his posterior odds are unlikely to change even by as much as a factor of 10, and he will remain favorable towards the null hypothesis. The same remarks hold true *mutatis mutandis* for the individual whose prior position strongly favors the alternative hypothesis. Thus, if we are willing to entertain the notion that the various users of these data are (perhaps unconsciously) applying some sort of Bayesian or quasi-Bayesian calculation, we can now understand why the strong rejection of the null hypothesis by classical tests has had little effect on people's opinions. These data are not decisive enough to change the minds of most people who already have even moderately strong opinions on the issue, and in fact even many neutral observers would probably regard the data as inconclusive. The evidence indicates that even people trained in statistics tend to change their personal odds much more conservatively than a Bayesian analysis would indicate (Lindley, 1970; Ed wards, 1982). So it should come as no surprise that data of the kind we are discussing have not changed many opinions on this subject.

Other Issues

Berger and Delampady (1987) point out that an exact point null hypothesis is rarely encountered in practice and is at best only an approximation to a more complicated actual situation. They remark,

> It is rare, and perhaps impossible, to have a null hypothesis that can be exactly modeled as $\vartheta = \vartheta_0$. One might feel that hypotheses such as
>
> H_0: A Subject has no ESP,
>
> or
>
> H_1: Talking to plants has no effect on their growth,
>
> are representable as exact (and believable) point nulls but, even here, minor biases in the experiments designed to test hypotheses will usually prevent exact representations as points.

Berger and Delampady discuss conditions under which it is appropriate to use the approximation of an exact point null. It is adequate when the expected bias in an experiment is substantially smaller than the sample standard deviation, say by a factor of at least 2. However, one who believes that the experiment might have a significant bias ought to use a prior that explicitly takes this possibility into account. Just as an example, someone who thinks it a realistic possibility that the experiment of Jahn *et. al.* was biased even at the level of 0.01% certainly ought to construct a prior on the null hypothesis that accounts for this expectation. This would convert the "no psi" hypothesis from a sharp point null hypothesis (i.e., a δ-function) into one that spreads the prior probability on the null hypothesis over a small interval of width ε centered on 0.5. This would be equivalent to regarding an effect of less than ε to be uninteresting as evidence for psi. In practice, such an assumption would eat into the prior probability that formerly was assigned to the alternative hypothesis, and the Bayes factor in favor of H would increase significantly. The calculations would also be more involved (see Berger and Delampady for details).

Shafer (1982) raises a different issue. He is bothered by the fact that the Bayesian analysis can strongly confirm the null hypothesis, whereas at the same time the sample

mean would be very precisely determined at a value that is significantly far away from the value implied by the null hypothesis. Shafer's example (a modified version of one presented by Lindley, 1977) considers the case of a burglary, during the commission of which a window is broken. A suspect is arrested, and a fragment of broken glass is retrieved from his clothes. Upon measurement, the index of refraction of the fragment is found to be several standard deviations different from the (very precisely determined) index of refraction of the window glass. The null hypothesis is that the two indices of refraction are the same; the alternative hypothesis (with a vague prior on the index of refraction of the sample) is that they are different. Despite the fact that the index of refraction of the fragment is several standard deviations different from that of the window, the null hypothesis is accepted with high posterior probability, because the prior distribution on the alternative hypothesis is broad, and so the integral over the alternative hypothesis is small compared to the integral over the null hypothesis. (This is the essence of the Jeffreys-Lindley paradox.) Shafer proposes an alternative methodology based on his theory of belief functions to account for what he views as shortcomings in the standard Bayesian approach.

Whatever the merits of Shafer's argument as applied to this example, there is an essential difference between it and the present case. The reason is that in his example, the glass fragment has a presumably stable index of refraction that can be measured consistently. If we were to remeasure it, we would expect to obtain a value that is within one or two standard deviations of the first measurement, and certainly not many standard deviations away. In parapsychology, on the other hand, the experience is quite different. We cannot expect that an observed effect will be stable from experiment to experiment, even if we repeat the same experiment, with the same subject, under the same conditions. On the contrary, the parapsychological literature is full of examples where a significant effect is observed, and then disappears (the "decline effect") and of examples where the size of the effect is otherwise very variable (e.g., the "shyness effect," the "linger effect," the "experimenter effect," etc.). Thus, I believe that a diffuse prior on the alternative hypothesis is appropriate in parapsychology.

The situation would be different if the psi hypothesis made precise predictions. In general, the bolder and more easily falsifiable are the predictions of a hypothesis, the more credibility it has when the observed effects are tolerably close to the predicted effects. For example, pre-Newtonian physics declared that there should be no correlation between the motions of heavenly and earthly bodies because physics was supposed to be very different in the two realms. Isaac Newton recognized that his hypothesis of universal gravitation required that the gravitational constant, as calculated from the motion of the Moon, must agree with that calculated from the fall of an apple, and indeed when he did the calculation he found that they agreed "pretty nearly." Despite the deficiencies in the data available to him, the surprising agreement between the two values surely must have encouraged Newton to think that he was on the right track, although he did not publish his results for many years and only after much prodding by his friend, Edmund Halley.

Lindley (1977) points out the importance of the "surprise value" of a prediction on the credibility of a hypothesis. He writes, "In the case of window glass there is considerable evidence about the distribution of refractive indices, some values being common, some rare. That such information is relevant is seen intuitively by considering the case where X

and *Y* are close together, being unusual indices. This gives greater evidence of identity than does the case where *X* and *Y* are equally close but are frequently occurring indices."

A priori, our "surprise" when we observe a value close to a sharp prediction is much greater than it would be if the theory made only a vague prediction. For example, consider a wholly imaginary world where stock market pundits provide standard deviations along with their predictions of market indexes. Suppose a pundit makes a prediction of the value of an index a year hence, and quotes a standard deviation of 1% for his prediction. We would probably be quite surprised if the actual value turned out to be within several percent of the prediction, and if this happened we might want to investigate the pundit more closely. By making a precise prediction, this pundit takes a great risk of being proven wrong (and losing our business). By the same token, when his prediction turns out even approximately correct, we are surprised, and the likelihood that we will follow his advice in the future may be increased. We would probably be less interested in a second pundit, who predicted the same value for the index as did the first, but who quoted a standard deviation of 20%. We would doubtless have little interest at all in a third pundit who informed us only that "the market will fluctuate," even though that prediction is virtually certain to be fulfilled!

This example illustrates the fact that scientific theories derive their power from the specificity of the predictions they make. The more difficult it is to adjust a theory to fit arbitrary data, and the more specific its predictions are, the more credible the theory becomes when the data do agree tolerably well with it. In a sense, an inflexible theory is "simpler" than one that is more accommodating, and by Occam's razor, more believable when both theories accommodate the data reasonably well. There is an intimate relationship between Occam's razor and Bayesian ideas, as Loredo (1990) has recently reaffirmed.

Finally, Casella and Berger (1987) describe a different kind of one-sided Bayesian test from the one I have used, and discuss its implications for Bayesian hypothesis testing. Their test gives results that are rather similar to the classical tests. Casella and Berger propose to test H: $z < a$ versus H': $z > a$. This approach would be very appropriate in certain drug testing situations, for example, where one is interested in whether a new drug is more effective (H') or less effective (H) than the standard treatment. If more effective, one would recommend that the new drug be adopted.

However, I believe that the test proposed by Casella and Berger is not appropriate to the present situation. One reason is that in parapsychology, large statistically significant negative results are generally considered to be positive evidence of paranormal effects. This phenomenon even has a name, "psi-missing." A test like Casella and Berger's would suggest considering a large, statistically significant negative result as evidence against paranormal effects, which seems odd. It is for this reason that I believe, as I stated above, that Jahn *et. al.* err in evaluating their data only with a one-tailed test. In their discussion they point out certain subjects who have "signatures" that tend to result in an excess of misses over hits, that is, psi-missing. Indeed, almost half (9/22) of their subjects had negative scores. Therefore, I believe that the two-tailed test is generally more appropriate to this kind of experiment and should usually be reported. Of course, the results of the one-tailed test can certainly be reported as well.

In parapsychology, the null hypothesis of negligible effect is distinguished from other hypotheses, since there is a genuine and nonnegligible belief that it might be true. As Berger and Sellke (1987) point out, under such circumstances we had better recognize the

distinguished nature of the hypothesis of "no effect" by explicitly reserving a significant amount of prior probability for it. Casella and Berger's proposed test does not satisfy this criterion. If one must use a one-sided test, then I think that the test that I proposed is more appropriate to parapsychology than theirs. In fact, we can broaden it to distinguish three mutually exclusive alternative hypotheses: H: $z = 0.5$, H': $z > 0.5$, and H'': $z < 0.5$. We can then calculate posterior probabilities for all three alternatives, that is, the null hypothesis, a significant result in the in tended direction. and a significant result in the contrary direction (psi-missing). In the present example using the data of Jahn *et. al.* (1987), the third hypothesis (psi-missing) is rejected pretty conclusively for most reasonable priors, a fact whose verification I leave as an exercise for the reader.

Conclusions

Bayesian analysis is very appropriate for parapsychological research. Parapsychology still remains "in limbo" with regard to mainstream science, and strong and even passionate opinions are evident on both sides of the issue. Therefore, it is especially appropriate to adopt a statistical methodology that both brings out into the open the (usually hidden) assumptions of the various players, and which also allows different people to reach different conclusions in an honest and aboveboard fashion that properly and publicly takes into account their personal attitudes and experiences. By making public the assumptions of the different players and therefore the way that they arrive at their conclusions, Bayesian analysis has the potential to help "clear the air" in this controversy. Its immunity to "stopping rule" problems can eliminate one objection that continues to be leveled by critics of parapsychology, perhaps more effectively than better protocols could ever do. It can also provide experimenters greater freedom to alter the protocol by taking additional data on promising subjects without compromising the statistical integrity of the experiment.

Parapsychologists need to face Berger and Delampady's (1987) point about bias squarely. It is simply wrong to think that one can obtain arbitrarily high accuracy by simply taking more and more data. No matter how carefully an experiment is designed, and no matter how diligently the protocol is followed, biases will inevitably dominate at some point, and experience in many fields shows that this point is generally reached sooner rather than later (Shafer, 1982, p.326). Utts (1988) points out the significance of this fact for parapsychology.

Unlike mainstream sciences, parapsychology suffers from a lack of critical experiments that produce large and unmistakable effects that can be duplicated frequently, if not consistently, under properly controlled conditions. Instead, parapsychology has come to rely on experiments like the one studied here, where the effects are very small and are only validated statistically. This research program has failed to convince skeptics, many of whom suspect that the "signal" consists largely of unrecognized experimental bias, even if they cannot explicitly identify its source. Glymore (1987) has urged parapsychologists to "go for the Big Stuff," and in my opinion this is good advice. As Edwards, Lindman, and Savage (1963) remark, "Rejection of a null hypothesis is best when it is interocular [hits you between the eyes]."

Navon (1987) remarks:

> Science would thus do well to consider the compatibility of a datum with prior *knowledge*. This maxim underlies the Bayesian approach to inference. Although Bayesian statistics in itself is fairly problematic, the main principle of Bayesian inference can hardly he disputed. The reason is that science should be (and, one hopes, is) a flexible but viscous system, one that can move from a local equilibrium state to a globally better one yet remain stable once it reaches an equilibrium. If it were not, we would be continually wavering among alternative hypotheses. Because vacillation is undesirable, but so too is prepossession, the Bayesian approach seems to represent a golden mean. Now, because the prior probability of the psi hypothesis, even merely as an empirical anomaly, is extremely low, applying the Bayesian rule to the data, questionable and undiagnostic as they already are, must be a *coup de grace* to that hypothesis.

Whether one agrees with Navon's final assessment or not, when viewed from a Bayesian perspective his remarks go a long way towards explaining why the evidence thus far produced by parapsychologists has failed to move critics very far towards a more favorable view of the psi hypothesis. Utts (1988) points out that because of the strong opinions held by various individuals, the Bayesian approach "makes particular sense for parapsychology," and goes on to remark that "Posterior probabilities in Bayesian analysis are a function of both the prior probabilities and the strength of the evidence; it may be informative to formalize these opinions and to see how much evidence would be needed to increase the posterior probability of a psi hypothesis to a nonnegligible level when the prior probability was close to zero." In this paper I have tried to show how this might be done in practice.

References

Abelson, P.H. (1978, July). A stepchild of science starts to win friends. *U.S. News and World Report*, p. 41.

Alcock, J.E. (1990). *Science and supernature*. Buffalo, NY: Prometheus.

Berger, J.O. (1985). *Statistical decision theory and Bayesian analysis (Second Ed.).* New York: Springer-Verlag.

Berger, J.O., & Berry, D.A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*, 159- 165.

Berger, J.O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317-352.

Berger, J.O., & Sellke. T. (1987). Testing of a point null hypothesis: The irreconcilability of significance levels and evidence. *Journal of the American Statistical Association, 82,* 112-122.

Casella, G., & Berger, R.L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association, 82,* 106-111.

Cicchetti, D.V. (1987). Differentiating between the statistical and substantive significance of ESP Phenomena: Delta, kappa, psi, phi, or it's not all Greek to me. *Behavioral and Brain Sciences, 10,* 577-581.

Edwards, W., Lindman, H., & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193-242. [Reprinted in *Robustness of Bayesian analysis* (J. B. Kadane, ed.). Amsterdam: North-Holland, 1984.]

Edwards, W. (1982). Conservatism in human information processing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 359-369). Cambridge: Cambridge University Press.

Feller, W. (1957). *An introduction to probability theory and its applications*. New York: John Wiley and Sons.

Glymore C. (1987). ESP and the big stuff. *Behavioral and Brain Sciences, 10*, 590.

Good, I.J. (1982). Comment. *Journal of the American Statistical Association, 77*, 342-344.

Hansel, C.E.M. (1989). *The search for psychic power*. Buffalo, NY: Prometheus Books. Chapter 16.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. La Salle, Illinois: Open Court Press.

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

Lee, P.M. (1989). *Bayesian statistics: An introduction*. New York: Oxford University Press.

Lindley, D.V. (1957). A statistical paradox. *Biometrika, 44*, 187-192.

Lindley, D.V. (1970). In D.L. Meyer & R.O. Collier, Jr. (Eds), *Bayesian statistics* (pp. 111-112). Itasca, Illinois: F. E. Peacock Publishers.

Lindley, D.V. (1977). A problem in forensic Science. *Biometrika, 64,* 207-213.

Loredo, T.J. (1990). From Laplace to Supernova SN 1987A: Bayesian inference in astrophysics. In P. Fougere (Ed.), *Maximum entropy and Bayesian methods*. (pp. 81-142). Dordrecht: Kluwer Academic Publishers (in press).

Navon, D. (1987). On rustles, wolf interpretations, and other wild speculations. *Behavioral and Brain Sciences, 10*, 599-601.

Palmer, J., & Rao, K.R. (1987).Where is the bias? *Behavioral and Brain Sciences, 10*, 618-627.

Press, S.J. (1989). *Bayesian Statistics: Principles, Models, and Applications*. New York: John Wiley & Sons.

Radin, D.I., & Nelson, R.D. (1989). Evidence for consciousness-related anomalies in random physical systems. *Foundations of Physics, 19*, 1499-1514.

Radin, D.I., May, E.C., & Thomas, M.J. (1986). Psi experiments with random number generators: Meta-analysis Part I . In D.H. Wiener & D.I. Radin (Eds.), *Research in parapsychology 1985* (pp. 14-17). Metuchen, NJ: Scarecrow Press.

Savage, L.J. (1972). *The Foundations of Statistics (2nd revised ed.).* New York: Dover Press.

Schmidt, H., & Pantas, L. (1972). Psi tests with internally different machines. *Journal of Parapsychology, 36,* 222-232.

Schmidt, H. (1969). Precognition of a quantum process. *Journal of Parapsychology, 33*, 99-108.

Schmidt, H. (1970). A PK test with electronic equipment. *Journal of Parapsychology. 34*, 176-181.

Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association*, 77, 325-351.

Smith, W.R., Dagle, E.F., Hill, M.D., & Mott-Smith, J. (1963). *Testing for extra sensory perception with a machine*. Data Sciences Laboratory Project 4610, Bedford, Massachusetts: Hanscom Field.

Tyrrell, G. N. M. (1936). Further research in extra-sensory perception. *Proceedings of the Society for Psychical Research*, *44,* 99-116.

Utts, J. (1988). Successful replication versus statistical significance. *Journal of Parapsychology, 52*, 305-320.

Zellner, A. (1971). *An introduction of Bayesian inference in econometrics*. New York: John Wiley & Sons.

Figures

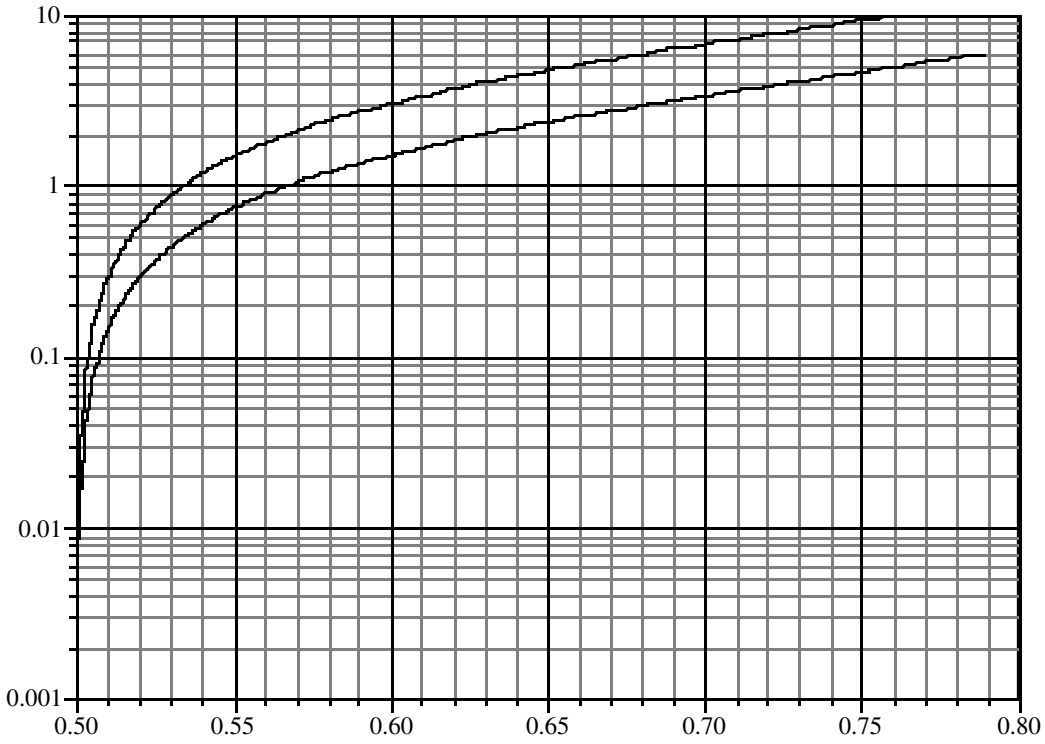

Fig. 1. The Bayes factor as a function of a. *Top curve*: Two-sided alternative. *Bottom curve*: One-sided alternative.

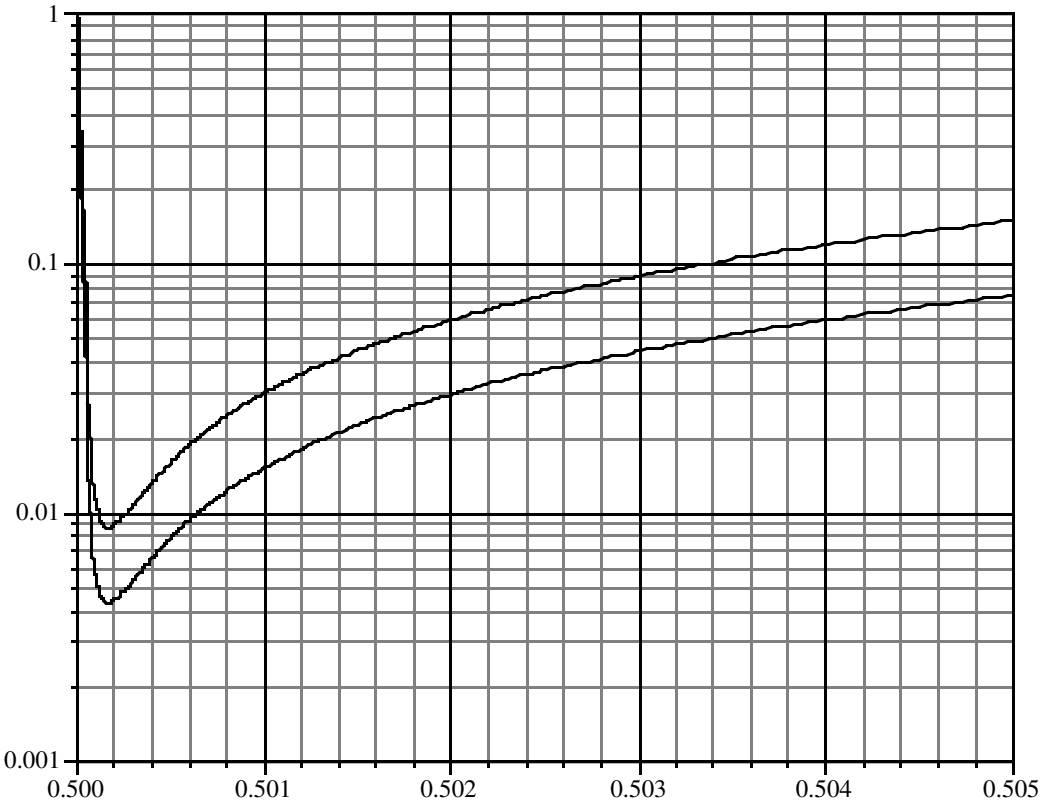

Fig. 2. Blowup of the region of Figure 1 near $a = 0.5$.

ERRATUM

Figures 1 and 2 of the paper, "Bayesian Analysis of Random Event Generator Data," which appeared in the Journal of Scientific Exploration, 4, 153-169, were drawn incorrectly. Corrected versions are shown in Figures 1 and 2 of this note. The effect of the change is to marginally strengthen the conclusions of the paper. In addition to correcting the figures, the following changes should be made in the body of the paper:

I) On p. 161, second paragraph, change "0.55" to "0.534" in both places. Change "0.5025 to "0.5017" and "0.4%" to "0.34%."

2) On p. 163, first paragraph, change "0.59" to "0.566" and change "0.505" to "0.5034."

I am indebted to Topher Cooper for discovering this error. I apologize to any who may have been inconvenienced by it.

William H. Jefferys
Department of Astronomy
University of Texas at Austin
Austin, TX 78712

[Note: These corrections have been included in this web-based version of the paper.]