Comment

M. J. Bayarri and James Berger

1. INTRODUCTION

There are many fascinating issues discussed in this paper. Several concern parapsychology itself and the interpretation of statistical methodology therein. We are not experts in parapsychology, and so have only one comment concerning such matters: In Section 3 we briefly discuss the need to switch from P-values to Bayes factors in discussing evidence concerning parapsychology.

A more general issue raised in the paper is that of replication. It is quite illuminating to consider the issue of replication from a Bayesian perspective, and this is done in Section 2 of our discussion.

2. REPLICATION

Many insightful observations concerning replication are given in the article, and these spurred us to determine if they could be quantified within Bayesian reasoning. Quantification requires clear delineation of the possible purposes of replication, and at least two are obvious. The first is simple reduction of random error, achieved by obtaining more observations from the replication. The second purpose is to search for possible bias in the original experiment. We use "bias" in a loose sense here, to refer to any of the huge number of ways in which the effects being measured by the experiment can differ from the actual effects of interest. Thus a clinical trial without a placebo can suffer a placebo "bias"; a survey can suffer a "bias" due to the sampling frame being unrepresentative of the actual population; and possible sources of bias in parapsychological experiments have been extensively discussed.

Replication to Reduce Random Error

If the sole goal of replication of an experiment is to reduce random error, matters are very straightforward. Reviewing the Bayesian way of studying this issue is, however, useful and will be done through the following simple example.

EXAMPLE 1. Consider the example from Tversky and Kahnemann (1982), in which an experiment results in a standardized test statistic of $z_1 = 2.46$. (We will assume normality to keep computations trivial.) The question is: What is the highest value of z_2 in a second set of data that would be considered a failure to replicate? Two possible precise versions of this question are: Question 1: What is the probability of observing z_2 for which the null hypothesis would be rejected in the replicated experiment? Question 2: What value of z_2 would leave one's overall opinion about the null hypothesis unchanged?

Consider the simple case where $Z_1 \sim N(z_1 \mid \theta, 1)$ and (independently) $Z_2 \sim N(z_2 \mid \theta, 1)$, where θ is the mean and 1 is the standard deviation of the normal distribution. Note that we are considering the case in which no experimental bias is suspected and so the means for each experiment are assumed to be the same.

Suppose that it is desired to test $H_0: \theta \le 0$ versus $H_1: \theta > 0$, and suppose that initial prior opinion about θ can be described by the noninformative prior $\pi(\theta) = 1$. We consider the one-sided testing problem with a constant prior in this section, because it is known that then the posterior probability of H_0, to be denoted by $P(H_0 \mid \text{data})$, equals the P-value, allowing us to avoid complications arising from differences between Bayesian and classical answers.

After observing $z_1 = 2.46$, the posterior distribution of θ is

$$\pi(\theta \mid z_1) = N(\theta \mid 2.46, 1).$$

Question 1 then has the answer (using predictive Bayesian reasoning)

$$\begin{aligned} &P(\text{rejecting at level } \alpha \mid z_1) \\ &\quad = \int_{c_\alpha}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z_2-\theta)^2} \pi(\theta \mid z_1)\, d\theta\, dz_2 \\ &\quad = 1 - \Phi\left(\frac{c_\alpha - 2.46}{\sqrt{2}}\right), \end{aligned}$$

where Φ is the standard normal cdf and c_α is the (one-sided) critical value corresponding to the level, α, of the test. For instance, if $\alpha = 0.05$, then this probability equals 0.7178, demonstrating that there is a quite substantial probability that the second experiment will fail to reject. If α is chosen to be the observed significance level from the first experiment, so that $c_\alpha = z_1$, then the probability that the

M. J. Bayarri is Titular Professor, Department of Statistics and Operations Research, University of Valencia, Avenida Dr. Moliner 50, 46100 Burjassot, Valencia, Spain. James Berger is the Richard M. Brumfield Distinguished Professor of Statistics, Purdue University, West Lafayette, Indiana 47907.

second experiment will reject is just 1/2. This is nothing but a statement of the well-known martingale property of Bayesianism, that what you "expect" to see in the future is just what you know today. In a sense, therefore, question 1 is exposed as being uninteresting.

Question 2 more properly focuses on the fact that the stated goal of replication here is simply to reduce uncertainty in stated conclusions. The answer to the question follows immediately from noting that the posterior from the combined data (z_1, z_2) is

$$\pi(\theta \mid z_1, z_2) = N(\theta \mid (z_1 + z_2)/2, 1/\sqrt{2}),$$

so that

$$P(H_0 \mid \text{data}) = \Phi(-(z_1 + z_2)/\sqrt{2}).$$

Setting this equal to $P(H_0 \mid z_1)$ and solving for z_2 yields $z_2 = (\sqrt{2} - 1)z_1 = 1.02$. Any value of z_2 greater than this will increase the total evidence against H_0, while any value smaller than 1.02 will decrease the evidence.

Replication to Detect Bias

The aspirin example dramatically raises the issue of bias detection as a motive for replication. Professor Utts observes that replication 1 gives results that are fully compatible with those of the original study, which could be interpreted as suggesting that there is no bias in the original study, while replication 2 would raise serious concerns of bias. We became very interested in the implicit suggestion that replication 2 would thus lead to less overall evidence against the null hypothesis than would replication 1, even though in isolation replication 2 was much more "significant" than was replication 1. In attempting to see if this is so, we considered the Bayesian approach to study of bias within the framework of the aspirin example.

EXAMPLE 2. For simplicity in the aspiring example, we reduce consideration to

θ = true difference in heart attack rates between aspirin and placebo populations multiplied by 1000;

Y = difference in observed heart attack rates between aspirin and placebo groups in original study multiplied by 1000;

X_i = difference in observed heart attack rates between aspirin and placebo groups in Replication i multiplied by 1000.

We assume that the replication studies are extremely well designed and implemented, so that one is very confident that the X_i have mean θ. Using normal approximations for convenience, the data can be summarized as

$$X_1 \sim N(x_1 \mid \theta, 4.82), \quad X_2 \sim N(x_2 \mid \theta, 3.63)$$

with actual observations $x_1 = 7.704$ and $x_2 = 13.07$.

Consider now the bias issue. We assume that the original experiment is somewhat suspect in this regard, and we will model bias by defining the mean of Y to be

$$\eta = \theta + \beta,$$

where β is the unknown bias. Then the data in the original experiment can be summarized by

$$Y \sim N(y \mid \eta, 1.54),$$

with the actual observation being $y = 7.707$.

Bayesian analysis requires specification of a prior distribution, $\pi(\beta)$, for the suspected amount of bias. Of particular interest then are the posterior distribution of β, assuming replication i has been performed, given by

$$\pi(\beta \mid y, x_i) \propto \pi(\beta)\exp\left\{-\frac{1}{2(1.54^2 + \sigma_i^2)}\,[\beta - (y - x_i)]^2\right\},$$

where σ_i^2 is the variance (4.82 or 3.63) from replication i; and the posterior probability of H_0, given by

$$P(H_0 \mid y, x_i) = \int_{-\infty}^{\infty} \Phi\left(-\frac{\sigma_i}{1.54\sqrt{\sigma_i^2 + 1.54^2}}\,(y - \beta) - \frac{1.54}{\sigma_i\sqrt{\sigma_i^2 + 1.54^2}}\,x_i\right)\pi(\beta \mid y, x_i)\,d\beta.$$

Recall that our goal here was to see if Bayesian analysis can reproduce the intuition that the original experiment could be trusted if replication 1 had been done, while it could not be trusted (in spite of its much larger sample size) had replication 2 been performed. Establishing this requires finding a prior distribution $\pi(\beta)$ for which $\pi(\beta \mid y, x_1)$ has little effect on $P(H_0 \mid y, x_1)$, but $\pi(\beta \mid y, x_2)$ has a large effect on $P(H_0 \mid y, x_2)$. To achieve the first objective, $\pi(\beta)$ must be tightly concentrated near zero. To achieve the second, $\pi(\beta)$ must be such that large $\mid y - x_2 \mid$, which suggests presence of a large bias, can result in a substantial shift of posterior mass for β away from zero.

A sensible candidate for the prior density $\pi(\beta)$ is the Cauchy $(0, V)$ density

$$\pi_V(\beta) = \frac{1}{\pi V\left[1 + (\theta / V)^2\right]}.$$

Flat-tailed densities, such as this, are well known to have the property that when discordant data is observed (e.g., when ($|y - x_2|$ is large), substantial mass shifts away from the prior center towards the likelihood center. It is easy to see that a normal prior for β can not have the desired behavior.

Our first surprise in consideration of these priors was how small V needed to be chosen in order for $P(H_0 \mid y,\ x_1)$ to be unaffected by the bias. For instance, even with $V = 1.54/100$ (recall that 1.54 was the standard deviation of Y from the original experiment), computation yields $P(H_0 \mid y,\ x_1) = 4.3 \times 10^{-5}$, compared with the P-value (and posterior probability from the original experiment assuming no bias) of 2.8×10^{-7}. There is a clear lesson here; even very small suspicions of bias can drastically alter a small P-value. Note that replication 1 is very consistent with the presence of no bias, and so the posterior distribution for the bias remains tightly concentrated near zero; for instance, the mean of the posterior for β is then 7.2×10^{-6}, and the standard deviation is 0.25.

When we turned attention to replication 2, we found that it did not seriously change the prior perceptions of bias. Examination quickly revealed the reason; even the maximum likelihood estimate of the bias is no more than 1.4 standard deviations from zero, which is not enough to change strong prior beliefs. We, therefore, considered a third experiment, defined in Table 1. Transforming to approximate normality, as before, yields

$$X_3 \sim N(x_3 \mid \theta, 3.48),$$

with $x_3 = 22.72$ being the actual observation. The maximum likelihood estimate of bias is now 3.95 standard deviations from zero, so there is potential for a substantial change in opinion about the bias.

Sure enough, computation when $V = 1.54/100$ yields that $E[\beta \mid y,\ x_3] = -4.9$ with (posterior) standard deviation equal to 6.62, which is a dramatic shift from prior opinion (that β is Cauchy (0, 1.54/100)). The effect of this is to essentially ignore the original experiment in overall assessments of evidence. For instance, $P(H_0 \mid y,\ x_3) = 3.81 \times 10^{-11}$, which is very close to $P(H_0 \mid x_3) = 3.29 \times 10^{-11}$. Note that, if β were set equal to zero, the overall posterior probability of H_0 (and P-value) would be 2.62×10^{-13}.

TABLE 1
Frequency of heart attacks in replication 3

	Yes	No
Aspirin	5	2309
Placebo	54	2116

Thus Bayesian reasoning can reproduce the intuition that replication which indicates bias can cast considerable doubt on the original experiment, while replication which provides no evidence of bias leaves evidence from the original experiment intact. Such behavior seems only obtainable, however, with flat-tailed priors for bias (such as the Cauchy) that are very concentrated (in comparison with the experimental standard deviation) near zero.

3. *P*-VALUES OR BAYES FACTORS?

Parapsychology experiments usually consider testing of H_0: No parapsychological effect exists. Such null hypotheses are often realistically represented as point nulls (see Berger and Delampady, 1987, for the reason that care must be taken in such representation), in which case it is known that there is a large difference between P-values and posterior probabilities (see Berger and Delampady, 1987, for review). The article by Jefferys (1990) dramatically illustrates this, showing that a very small P-value can actually correspond to evidence for H_0 when considered from a Bayesian perspective. (This is very related to the famous "Jeffreys" paradox.) The argument in favor of the Bayesian approach here is very strong, since it can be shown that the conflict holds for virtually any sensible prior distribution; a Bayesian answer can be wrong if the prior information turns out to be inaccurate, but a Bayesian answer that holds for all sensible priors is unassailable.

Since P-values simply cannot be viewed as meaningful in these situations, we found it of interest to reconsider the example in Section 5 from a Bayes factor perspective. We considered only analysis of the overall totals, that is, $x = 122$ successes out of $n = 355$ trials. Assuming a simple Bernoulli trial model with success probability θ, the goal is to test $H_0{:}\theta = 1/4$ versus $H_1{:}\theta \neq 1/4$.

To determine the Bayes factor here, one must specify $g(\theta)$, the conditional prior density on H_1. Consider choosing g to be uniform and symmetric, that is,

$$G_r(\theta) = \begin{cases} \dfrac{1}{2r}, & \text{for } \dfrac{1}{4} - r \leq \theta \leq \dfrac{1}{4} + r, \\ 0, & \text{otherwise}. \end{cases}$$

Crudely, r could be considered to be the maximum change in success probability that one would expect given that ESP exists. Also, these distributions are the "extreme points" over the class of symmetric unimodal conditional densities, so answers that hold over this class are also representative of answers over a much larger class. Note that here $r \leq 0.25$ (because $0 \leq \theta \leq 1$); for the given data the $\theta > 0.5$ are essentially irrelevant, but if it were deemed important to take them into account one could use the more sophisticated binomial analysis in Berger and Delampady (1987).

For g_r, the Bayes factor of H_1 to H_0, which is to be interpreted as the relative odds for the hypotheses provided by the data, is given by

$$B(r) = \frac{(1/(2r)) \int_{.25-r}^{.25+r} \theta^{122}(1-\theta)^{355-122}\, d\theta}{(1/4)^{122}(1-1/4)^{355-122}}$$

$$\cong \frac{1}{2r}(63.13) \cdot \left[\Phi\left(\frac{r-.0937}{.0252}\right) + \Phi\left(\frac{-(r+.0937)}{.0252}\right)\right].$$

This is graphed in Figure 1.

The P-value for this problem was 0.00005, indicating overwhelming evidence against H_0 from a classical perspective. In contrast to the situation studied by Jefferys (1990), the Bayes factor here does not completely reverse the conclusion, showing that there are very reasonable values of r for which the evidence against H_0 is moderately strong, for example 100/1 or 200/1. Of course, this evidence is probably not of sufficient strength to overcome strong prior opinions against H_0 (one obtains final posterior odds by multiplying prior odds by the Bayes factor). To properly assess strength of evidence, we feel that such Bayes factor computations should become standard in parapsychology.

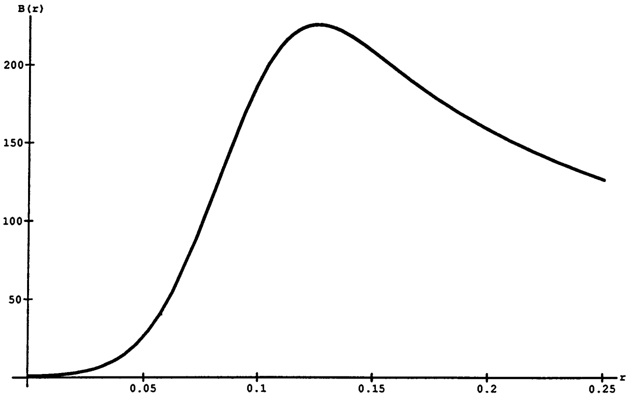

FIG. 1. *The Bayes factor of H_1 to H_0 as a function of r, the maximum change in success probability that is expected given that ESP exists, for the ganzfeld experiment.*

As mentioned by Professor Utts, Bayesian methods have additional potential in situations such as this, by allowing unrealistic models of iid trials to be replaced by hierarchical models reflecting differing abilities among subjects.

ACKNOWLEDGMENTS

M. J. Bayarri's research was supported in part by the Spanish Ministry of Education and Science under DGICYT Grant BE91-038, while visiting Purdue University. James Berger's research was supported by NSF Grant DMS-89-23071.

Comment

Ree Dawson

This paper offers readers interested in statistical science multiple views of the controversial history of parapsychology and how statistics has contributed to its development. It first provides an account of how both design and inferential aspects of statistics have been pivotal issues in evaluating the outcomes of experiments that study psi abilities. It then emphasizes how the idea of science as replication has been key in this field in which results have not been conclusive or consistent and thus meta-analysis has been at the heart of the literature in parapsychology. The author not only reviews past debate on how to interpret repeated psi studies, but also provides very detailed information on the Honorton–Hyman argument, a nice illustration of the challenges of resolving such de-

Ree Dawson is Senior Statistician, New England Biomedical Research Foundation, and Statistical Consultant, RFE/RL Research Institute. Her mailing address is 177 Morrison Avenue, Somerville, Massachusetts 02144.