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.

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.

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 $H_0\colon \theta \le 0$ versus $H_1\colon \theta > 0$, 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)

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

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