Journal of Scientific Exploration, Vol. *6,* No. 1, pp. 23–45, 1992 0892-3310/92

On the Bayesian Analysis of REG Data

YORK H. DOBYNS
Princeton Engineering Anomalies Research
Princeton University, Princeton, NJ 08544

Abstract—Bayesian analysis may profitably be applied to anomalous data obtained in Random Event Generator and similar human/machine experiments, but only by proceeding from sensible prior probability estimates. Unreasonable estimates or strongly conflicting initial hypotheses can project the analysis into contradictory and misleading results. Depending upon the choice of prior and other factors, the results of Bayesian analysis range from confirmation of classical analysis to complete disagreement, and for this reason classical estimates seem more reliable for the interpretation of data of this class.

Introduction

The relative validity of Bayesian versus classical statistics is an ongoing argument in the statistical community. The Princeton Engineering Anomalies Research program (PEAR) has heretofore used classical statistics exclusively in its published results, as a matter of conscious policy, on the grounds that the explicit inclusion of prior probability estimates in the analysis might divert meaningful discussion of experimental biases and controls into debate over the suitability of various priors. Nonetheless, Bayesian analysis can offer some clarifications, particularly in discriminating evidence from prior belief, and is therefore worth examination.

In this article we apply the Bayesian statistical approach to a large body of random event generator (REG) data acquired over an eight-year period of experimentation in the human/machine interaction portion of the Princeton Engineering Anomalies Research (PEAR) program. When assessed by classical statistical tests, these data display strong, reproducible, operator-specific anomalies, clearly correlated with pre-recorded intention and, to a lesser degree, with a variety of secondary experimental parameters (Nelson, Dunne, and Jahn, 1984). When assessed by a Bayesian formalism, the conclusions can range from a confirmation of the results of the classical analysis with essentially the same p-value against the null hypothesis, to a confirmation of the null hypothesis at 12 to 1 odds against a suitably chosen alternate, depending on how the analysis is done and what hypotheses are chosen for comparison.

The intent of this paper is to examine both the range of conclusions possible from Bayesian analysis as applied to a specific data set and the implications

of that range. The empirical meaning of families of prior probabilities that lead to related conclusions is of particular interest. We will also examine the relation between Bayesian odds adjustments and classical p-values in light of considerations of statistical power and the likelihoods of what a classical statistician would call "Type I" and "Type II" error. (At various times it will be necessary to contrast Bayesian with non-Bayesian approaches which are variously called "frequentist", "Fisherian", or "sampling theory" statistics. While acknowledging that non-Bayesian statistics are a conglomerate category on the order of "nonelephant animals," for purposes of this discussion any analytical approach that does not include an explicit role for subjective prior probabilities will be called "classical.")

Elementary Bayesian Analysis

Bayes' theorem ($p(\theta|y)p(y) = p(y|\theta)p(\theta)$) is a fundamental result of probability theory from the properties of contingent probabilities. Its analytical application is that of revising prior probability estimates in light of evidence, that is, of determining the extent to which various possibilities are supported by a given empirical outcome. In the most basic application of Bayesian analysis one is assumed to have a model or hypothesis with one or more adjustable parameters 8. One's state of prior knowledge, or ignorance, may be expressed as a prior probability distribution $\pi_0(\theta)$ over the space of possible 8 values. It is further assumed that some method exists by which a probability $\pi(y|\theta)$ can be computed for any possible experimental outcome y, given a definite parameter value θ. In standard REG experiments, for example, 8 consists of a single parameter, the probability p with which the machine generates a hit in an elementary Bernoulli trial. The probability of exactly s successes in a set of n trials is then given by the binomial formula

$$\pi(n, s|p) = \binom{n}{s} p^s(1 - p)^{n-s} \qquad (1$$

where $\binom{n}{s}$ is the combination of n elements taken s at a time, namely $n!/[s!(n - s)!]$.

By Bayes' theorem, Equation (1) may alternately be regarded as the probability of a parameter value p, given actual data n and s. When cast in this form, π is more commonly called the likelihood, ℓ. The general recipe for updating one's knowledge of parameter(s) 8 in light of evidence y is expressed by

$$\pi_1(\theta|y) \propto \ell(\theta|y)\pi_0(\theta) \qquad (2)$$

where $\pi_1(\theta|y)$ is the posterior probability distribution among possible values for 8, given the prior distribution π_0 and the likelihood ℓ. For the case of REG data,

$$\pi_1(p \mid n, s) \propto \ell(p \mid n, s)\pi_0(p) = \binom{n}{s} p^s(1 - p)^{n-s}\pi_0(p). \quad (3)$$

Note that (2) and (3) are expressed as proportionalities rather than equalities. While there is always a normalization for ℓ such that $Q(6|y)\pi_0(\theta)$ has a total integral of 1, this normalization in general depends on $\pi_0(\theta)$. Specifically, the use of $L(6|y) = Q(6|y)/\int_\theta d\theta \ell(\theta|y)\pi_0(\theta)$, where $\int_\theta$ denotes integration over all possible values of 6, produces the correct normalization. The relations (2) and (3), on the other hand, express Q purely as a function of y and θ without reference to prior probabilities. The conceptual clarity of stepping from one state of knowledge about 6, or probability distribution over θ, to another, with the aid of a quantity that depends only on objective evidence, thus entails the cost of renormalizing the posterior probabilities to a total strength of 1 as the last step in the calculation.

Since it is already a proportionality, expression (3) may be simplified to

$$\pi_1(p \mid n, s) \propto p^s(1 - p)^{n-s}\pi_0(p) \quad (4)$$

where the combinatorial factor, lacking p dependence, has been subsumed into the normalization. This form illustrates an important feature of Bayesian analysis. After prior probabilities have been adjusted in the light of first evidence, the resulting posterior probabilities may then be used as priors for subsequent calculation based on new evidence. For example, after two stages of such iteration,

$$\pi_2(\theta \mid y_1, y_2) \propto \ell(\theta \mid y_2)\pi_1(\theta \mid y_1) \propto \ell(\theta \mid y_2)\ell(\theta \mid y_1)\pi_0(\theta). \quad (5)$$

Note, however, that one can argue with equal merit that the posterior probabilities after both data sets y_1, y_2 are available should just be $\pi_2(\theta|y_1, y_2) \propto \ell(\theta|y_1 + y_2)\pi_0(\theta)$. For these formulas to produce different values of π_2 would be contradictory, i.e., the evidence would support different conclusions depending on the sequence in which it was evaluated. This can be avoided only if ℓ has the property $Q(6|y_1 + y_2) \propto Q(6|y_1)\ell(\theta|y_2)$. The likelihood function $\ell = p^s(1 - p)^{n-s}$ indeed has this essential addition property:

$$\ell(p \mid n_1 + n_2, s_1 + s_2) = p^{s_1+s_2}(1 - p)^{(n_1+n_2)-(s_1+s_2)} = \ell(p \mid n_2, s_2)\ell(p \mid n_1, s_1). \quad (6)$$

Since Bayesian formalism is internally consistent, a calculation such as (6) is essentially a check on the validity of the likelihood function; any legitimate ℓ must have the appropriate addition property.

Bayesian Analysis of REG Data

Let us now apply Bayesian formalism to the body of REG data presented in Margins of Reality (Jahn and Dunne, 1987), also published in the Journal for Scientific Exploration (Jahn, Dunne, and Nelson, 1987). From the summary

table on pp. 352–53 of the book and pp. 30–32 of the Journal article, the "**APK** data consist of some 522,450 experimental trials totalling n = 104,490,000 binary Bernoulli events. The Z score of 3.614 is tantamount to an excess of 18,471 hits in the direction of intention. From relation (4), the likelihood function for Bernoulli trials is $\ell(p \mid n, s) = p^s(1 - p)^{n-s}$. The mean of this p distribution is $(s + 1)/(n + 2)$; for large n its standard deviation is a $\sigma = \frac{1}{2\sqrt{n}} + O(1/n)$; and it becomes essentially normal. With the values above,

$$\ell(p \mid n, s) = p^{52263471}(1 - p)^{52226529}; \quad \mu_\ell = 0.5001768; \quad \sigma_\ell = 4.89 \times 10^{-5}. \quad (7)$$

Figure 1 shows this likelihood function for the REG data against a range of p values from 0.4999 to 0.5004. Also shown for comparison is the interval of p values that are credible under the null hypothesis, calculated as outlined below.

The REG device itself is constructed to generate groups of Bernoulli trials with an underlying hit probability as close to exactly 0.5 as possible (Nelson, Bradish, and Dobyns, 1989). As part of its internal processing, it compares the random string of positive and negative logic pulses from the noise generator with a sequence of regularly alternating positive and negative pulses generated by a separate circuit. Matches with this alternating "template" are then reported as hits in the final sum. This technique effectively cancels any systematic bias in the raw noise process. While it is conceivable that some bias in the final output could still occur, if for example some part of the apparatus contributed an oscillatory component to the noise that happened to be exactly in phase with the template, or if the counting module should systematically malfunction, such remote possibilitiesare precluded by a number of internal failsafes and counterchecks incorporated in the circuitry. The device was extensively and regularly calibrated during the period that the Margins data were collected, and from these calibration data it was established that, if p is expressed as 0.5 + 6, then $|\delta| < 0.0002$, with no lower bound established.

Yet further protection against bias is provided by the experimental protocol, wherein each operator generates approximately equal amounts of data in three experimental conditions. These are labeled "high," "low," and "baseline" in accordance with the operator's pre-recorded intentions. The "**APK**" data in Margins are differential combinations of "high" and "low" intention data; the combined result is equivalent to inverting the definition of success for the "low" data and computing the deviation of the resulting composite sequence of high and low efforts from chance expectation. Thus, to survive in the **APK** results, any residual artifactual bias of the device or the data processing would itself have to correlate with the operator's intention. Specifically, if $p_o = 0.5$ + 6 is the probability of an "on" bit, a data set containing a total of N_h bits from the high intention and N_l bits from the low intention (where the goal is to get "off' bits) will have a null-hypothesis p in the **APK** of:

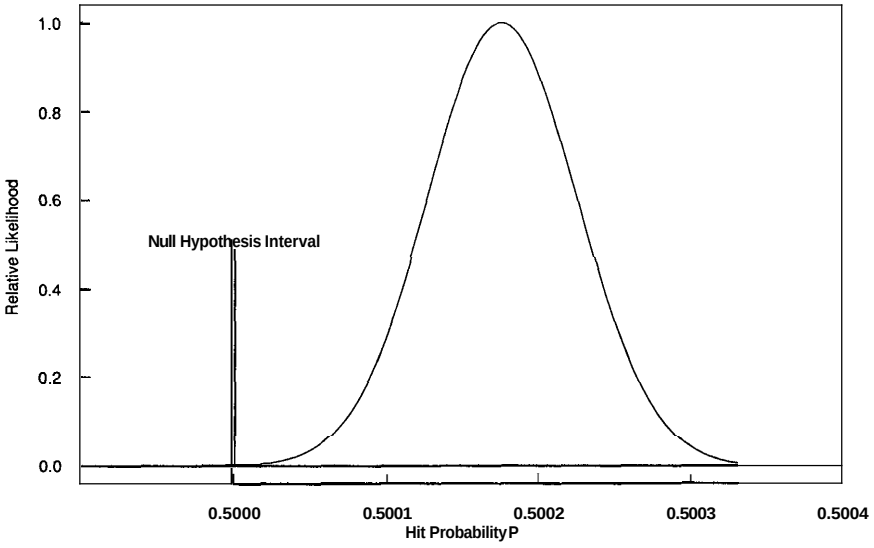

Fig. 1. **Likelihood function for PEAR REG data.**

$$p_\Delta = \frac{N_h p_o + N_l(1 - p_o)}{N_h + N_l} = 0.5 + \delta\frac{N_h - N_l}{N_h + N_l}. \tag{8}$$

Thus, when $N_h = N_l$, $p_\Delta \equiv 0.5$, regardless of the value of 6. For the actual ***Margins*** data, with $N_h = 52{,}530{,}000$ bits and $N_l = 51{,}960{,}000$ bits, $(N_h - N_l)/(N_h + N_l) = 0.0055$. Given $|6| < 0.0002$ as above, the maximum possible artifactual deviation from $p\Delta = 0.5$ is 1.1×10^{-6}. This value is the source of the null hypothesis interval shown in Figure 1.

While the issue of possible sources of bias in the REG data could be treated at considerably greater length (see, for a fuller treatment, Nelson ***et al.,*** 1989) such discussion is a separate issue from the statistical interpretation of the data. It has been mentioned here only to explain the derivation of the null hypothesis interval.

Having established the likelihood function **(Eq.** 7), let us now consider various sets of prior probabilities with which ℓ may be combined to arrive at a posterior probability distribution for the value of p in the actual experiment. First, consider the prior probability corresponding to extreme skepticism. **A** person who regards any influence of consciousness on the REG output to be impossible a ***priori*** should, by the tenets of Bayesian analysis, choose a prior $\pi_0(p) = \delta(p - 0.5)$, where $\delta(x)$ is the standard Dirac delta function defined by the property $\int_a^b f(x)\delta(x - x_0)dx \equiv f(x_0)$ for any f and any a, b such that $a < x_0 < b$. It then follows that $\pi_1(p)$ will also be a delta function, and after normalizing must in fact be the same function. Since this

choice of prior probability is clearly impervious to any conceivable evidence, it is illegitimate in any effort to learn from new information, however strongly held on philosophical grounds.

As an extreme alternative, one might select a prior evincing complete ignorance as to the value of p, by regarding all the possible values of p as equally probable: $\pi_0(p) = 1$ for $p \in [0, 1]$ as illustrated in Figure 2a. With this prior the posterior probability $\pi_1(p)$ must, of course, have exactly the same shape as ℓ. This replicates classical analysis in the following sense: ℓ is normal with its center 3.614 standard deviations away from p = 0.5, so, if we define confidence intervals in p centered on the region of maximum posterior probability, we may include as much as 99.97% of $\pi_1(p)$ before the interval becomes compatible with a point null hypothesis, corresponding to the 3.0×10^{-4} p value of a two-tailed classical test. Accounting for the actual spread of the null hypothesis slightly narrows this interval, raising the equivalent p value to 3.3×10^{-4}.

It is, however, unnecessary to assume this level of ignorance to arrive at a very similar result. For example, one might regard it as plausible, in light of the measures taken to force $p \approx 0.5$, that p ought to have some value in a narrow range centered about 0.5 but that within that range there is no strong reason to prefer one p over another. This defines a one-parameter family of "rectangular" priors characterized by their width w:

$$\pi_0(w, p) = \begin{cases} 1/w, & \text{if } p \in [\frac{1-w}{2}, \frac{1+w}{2}]; \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Figure 2b illustrates the member of this family with $w = 10^{-3}$. Use of this prior essentially replicates the result from the uniform prior of Figure 2a, since it still includes all of the likelihood function except for tiny contributions in the extreme tails. In consequence, π_1 has the same shape as in the previous case for the region $0.5 - \frac{w}{2} < p < 0.5 + \frac{w}{2}$, but is uniformly augmented by a multiplicative factor to compensate for the missing tails. Until w is made small enough that $0.5 + \frac{w}{2}$ comes within a few standard deviations of the maximum of ℓ the effects of this correction remain negligible.

Obviously, if the prior is made sufficiently narrow it will become indistinguishable from the null hypothesis interval and the resulting posterior probability can no longer exclude the null hypothesis interval from the region of high likelihood. Figure 3 displays the equivalent p value with which the null hypothesis is excluded for a range of widths of the prior; as above, this p is the conjugate probability to the widest confidence interval about the maximum of π_1 that does not include any of the null hypothesis interval. The line labeled "Breakpoint" marks a value of special interest for the width of the prior. For wider priors, the upper limit of the confidence range for the Bernoulli p is established by the symmetry condition about the peak, and the condition that the interval not include the null hypothesis range. For narrower priors, this upper limit is dictated by the width of the prior itself. It is unsurprising

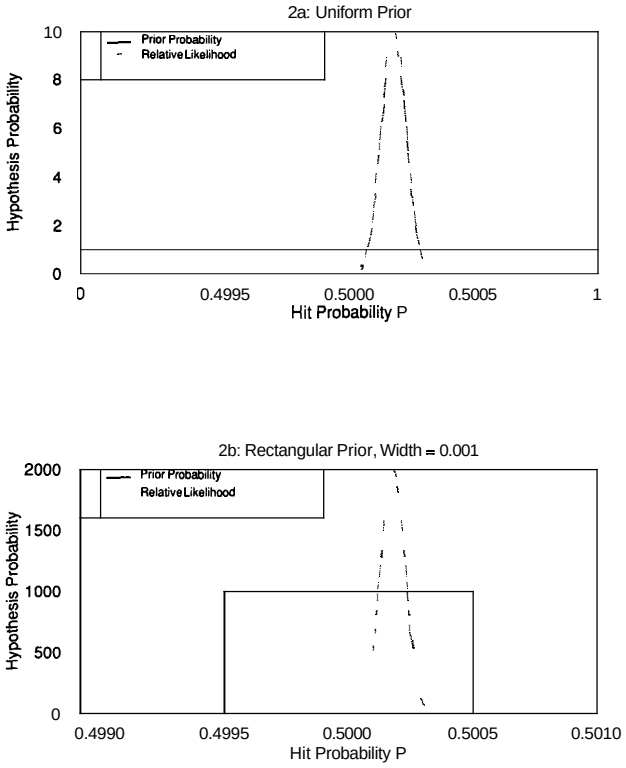

Fig. 2. Likelihood and different priors.

that this change of regimes is accompanied by an inflection in the p value of the null hypothesis. Beyond the left edge of Figure 3, we should note that when the width of the prior drops to 2.2×10^{-6}, the same as the null hypothesis, the ***p*** value of the null rises to *1.* This is essentially the same imperviousness previously seen in the delta-function prior. Indeed, the family of rectangular priors tends toward a delta function in the limit as the width goes to zero. However, values consistent with the null hypothesis are still excluded at p = 0.05 for *w* as small as 8.7×10^{-5}. Note that this is of the same order as the width of the likelihood function itself.

Further perspective on the interplay of the evidence with a prior preference for the null hypothesis interval may be obtained by consideringanother family of priors that specifically favor the null hypothesis to varying degrees but do not have sharp cutoffs of probability. Let $\pi_0(k, \mathrm{p}) = [(2k + 1)!/(k!)^2]p^k(1 - p)^k$ for any k. All of these functions are properly normalized probability distributions, with mean 0.5 and standard deviation $a = \tfrac{1}{2}\sqrt{(k - 1)/(2k^2 + 5k + 3)}$, which tends to $\dfrac{1}{2\sqrt{2k}}$ for large ***k.*** These functions also become increasingly normal for greater k. As in the previous case, they tend to a delta function in the limit $k \to \infty$. When one of these functions is used as a prior with ℓ from

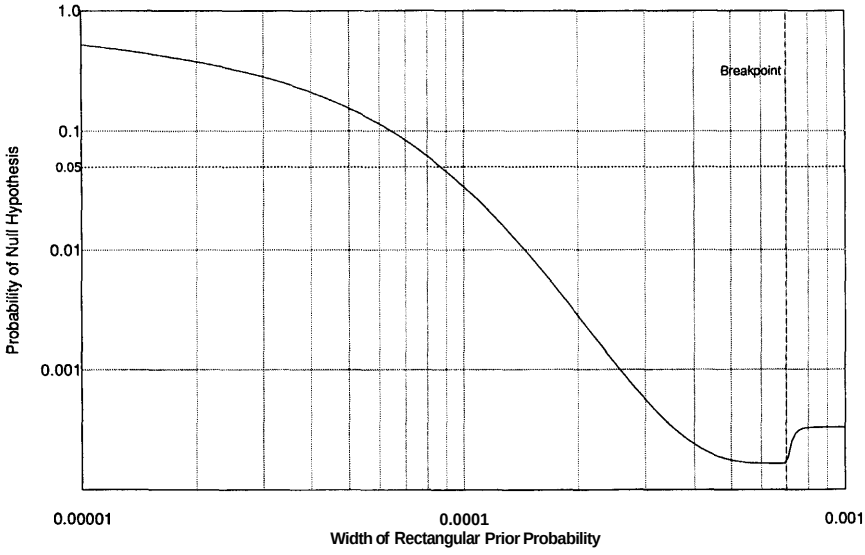

Fig. 3. Conjugate confidence intervals from **rectangular priors.**

the Margins data, the resulting π_1 has mean $\mu_1 = (s + 1 + k)/(n + 2 + 2k)$ and standard deviation $a_1 = \frac{1}{2}\sqrt{n + 2k}$ (in the large-n approximation, which is clearly justified), as can be seen from the functional form of Q and the fact that multiplying by π_0 is equivalent to the substitution $s \to s + k$, $n \to n + 2k$, up to normalization. The equivalent Z score, that is, the number of its own standard deviations that separate the peak of the posterior probability distribution from the null hypothesis, becomes $Z = (2s - n)/\sqrt{n + 2k}$. While this clearly tends toward zero as $k \to \infty$, it is also clear that large values of k, and hence extremely narrow priors, are needed to change the result appreciably. Figure 4 presents the equivalent p value, as defined for Figure **3,** for this family of priors as a function of k. Also shown is the width (standard deviation) of the prior, indicative of how strongly the null hypothesis is favored. Note that to drive the p value above 0.05 (that is, to bring the null hypothesis interval within the 95% confidence interval of the posterior probability) a $k > 10^8$, or a $a < 3 \times 10^{-5}$, is required. Here the characteristicscale of the prior is actually narrower than that of the likelihood.

An alternative way of favoring a narrowly defined region of probability often employed in Bayesian analysis, as pointed out by the reviewer of an earlier version of this work, is to put some of the prior probability in a "lump" at the preferred value. In this case, for example, one might modify any of the priors above by multiplying it by $1 - a$ and then adding $a\delta(p - 0.5)$, for any $0 < a < 1$; this inflates the degree of probability accorded the null hypothesis. Large values of a are not very interesting, since $a = 1$ replicates the completely impervious delta-function distribution. Consider a family of priors that might be regarded as plausible by an analyst who believes the null hypothesis has

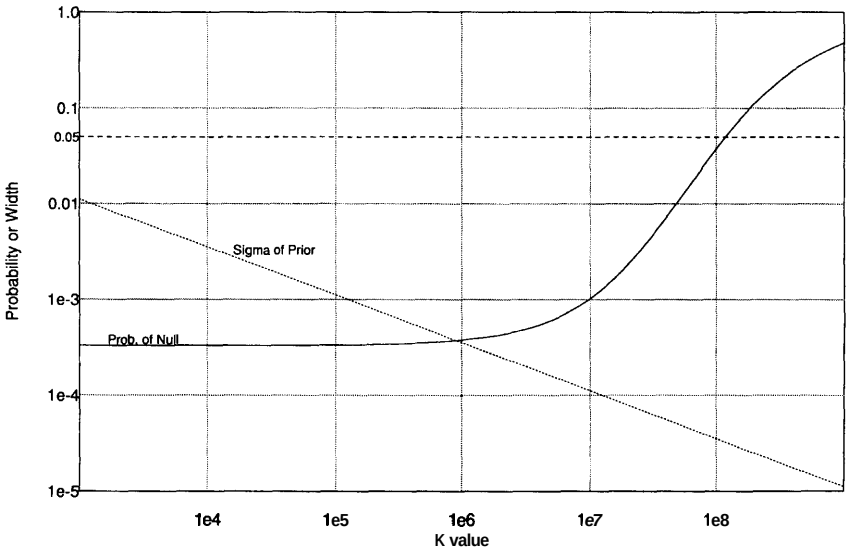

Fig. 4. Conjugate probabilities from K-family of priors.

considerable support but who has no reason to prefer one value of p over another within some reasonable range for the alternate. This might be represented as $r_0(\sim w, p) = a\delta(p - 0.5) + (1 - a)\pi_w(p)$, where $\pi_w(p)$ is the same "rectangular" prior defined as $\pi_0(w, p)$ in Eq. 9. Thus $\pi_0(a, w, p)$ is a two-parameter family of priors in a, the extra weight initially assigned to the null hypothesis, and w, the range of plausible alternatives. The confidence-interval formulation discussed above is somewhat awkward for the posterior probability resulting from these functions, since they are highly bimodal. However, this bimodality arises from the preservation of the delta-function component and also suggests that the posterior probability of the null hypothesis, given this prior, may be computed from the strength of the delta-function null in the (normalized) posterior probability π_1. The contribution from the part of the π_w component compatible with the null hypothesis is negligible for most values of a and w.

The upper portion of Figure 5 presents a contour plot of the posterior probability of $p = 0.5$ for a range of a and w values. Both scales are logarithmic, with grid lines shown at 1, 3, 5, 7, and 9 times even powers of 10. For a values as large as 0.9, the posterior probability of the null is less than 0.05 for $w \approx 5 \times 10^{-4}$. As a grows the calculation becomes less sensitive to w and less responsive to the data, as expected.

The lower portion of Figure 5 shows a related quantity of interest, the relative strength of the null hypothesis in the prior and posterior distributions as given by the coefficient of the delta-function component. This can be regarded as the degree to which the null hypothesis component is amplified by the evidence. Two noteworthy features are that for small a values this

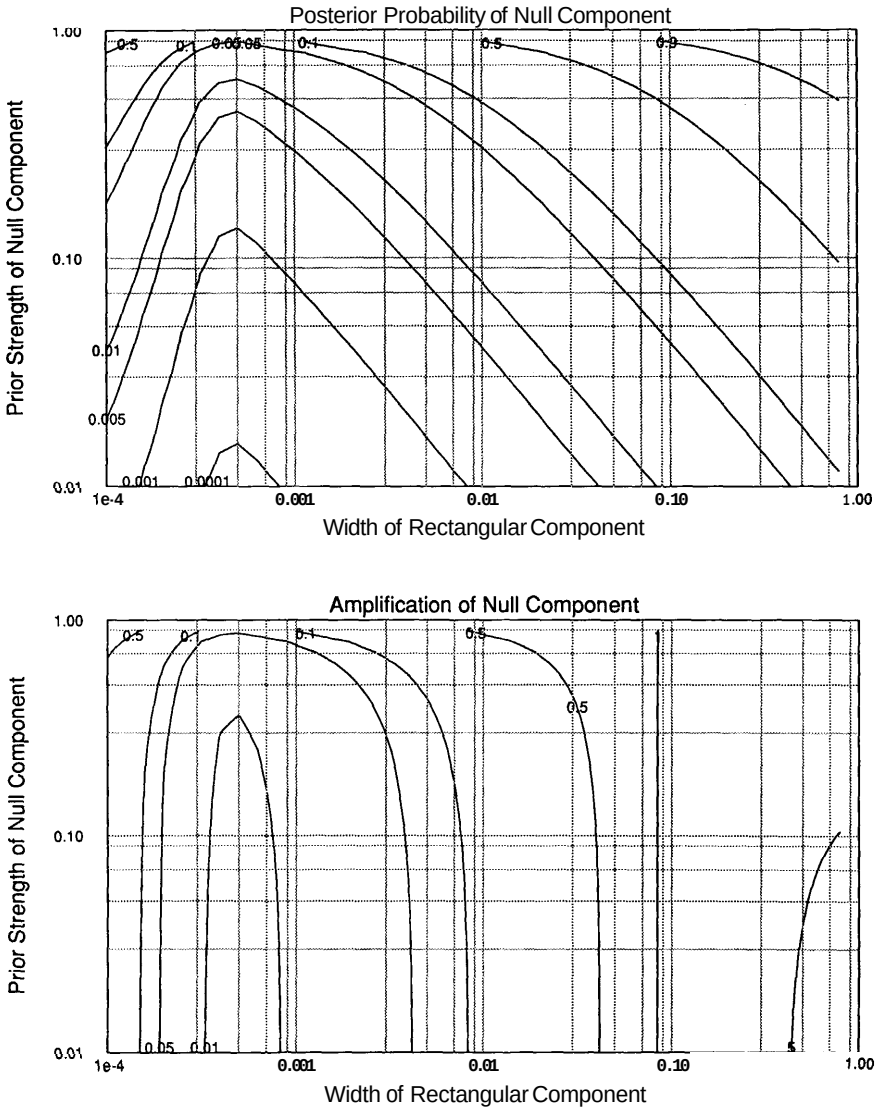

Fig. *5.*

amplification factor tends toward a constant depending only on w, and that even for a ≈ 0.85, the null hypothesis emerges twenty times less likely after accounting for the evidence for $w = 5 \times 10^{-4}$.

In summary, an examination of various possible prior probability distributions leads to conclusions ranging from confirmation of the classical odds against the null hypothesis to confirmation of the null hypothesis, depending on one's choice of prior. Priors that lead to confirmation, or low odds against, the null hypothesis, are associated with large concentrations of probability on

the null hypothesis, or ranges around the null that are narrow compared to the likelihood function. In other words, they must be relatively impervious to evidence.

For all of these examples, the evidence (as manifested in the likelihood function) has remained constant. The variability of the conclusions has resulted entirely from the various choices of prior probability distribution. With the pure delta-function prior standing as a cautionary example of a prior belief that cannot be shaken by any evidence whatsoever, it seems suggestive that those priors which lead to conclusions most strongly in disagreement with the classical analysis are precisely those which most nearly approach the delta-function. A possibly oversimplified summation is that the likelihood function, taken alone, would lead to the same conclusion as a classical analysis, while the more an analyst wishes to favor the null hypothesis a priori, the more the posterior conclusions will likewise favor the null. This at least suggests that a prior hypothesis leading to strong disagreement with classical analysis may be inappropriate to a given problem.

Concerns of appropriate choices of prior hypotheses will be addressed further below, in light of another method of analysis.

Bayesian Hypothesis Testing and Lindley's Paradox

The last example in the previous section was chosen in part because it leads rather directly to the question of using Bayesian analysis to compare two distinct hypotheses, rather than evaluating a parameter range under a single hypothesis. Consider for example the hypotheses $\pi_0(\theta)$ and $\pi_1(\theta)$, where π_1 now denotes an alternative prior. Let p_0 and p_1 denote prior probabilities on the hypotheses, with $p_0 + p_1 = 1$ so that the two hypotheses comprise exhaustive alternatives. The relative likelihood of the hypotheses can also be stated as the prior odds $\Omega = p_0/p_1$.

Given the two hypotheses and their respective prior probabilities, an overall prior probability distribution for θ can be constructed $\pi(\theta) = p_0\pi_0(\theta) + p_1\pi_1(\theta)$. This may then be used in a Bayesian calculation resulting in a posterior probability $\pi'(\theta) = L(\theta|y)\pi(\theta)$, where $L(\theta|y) \equiv \ell(\theta|y)/\int_\theta \ell(\theta|y)\pi(\theta)d\theta$ is the normalized likelihood. This posterior probability can unambiguously be divided into components arising from the two hypotheses, $\pi' = \pi'_0 + \pi'_1$, such that

$$\pi'_0(\theta|y = L(\theta|y)p_0\pi_0(\theta) \quad \text{and} \quad \pi'_1(\theta|y) = L(\theta|y)p_1\pi_1(\theta). \tag{10}$$

The posterior probabilities for the two hypotheses are clearly the total integrals of their respective contributions to the overall posterior probability: $p'_0 = \int_\theta \pi'_0(\theta|y)d\theta$ and likewise for p'_1. Thus the posterior odds are

$$\Omega'(\theta \mid y) = \frac{p_0'}{p_1'} = \frac{\int_\theta L(\theta \mid y) p_0 \pi_0(\theta) d\theta}{\int_\theta L(\theta \mid y) p_1 \pi_1(\theta) d\theta}$$
$$= \left(\frac{\int_\theta \mathcal{Q}(\theta \mid y) \pi_0(\theta) d\theta}{\int_\theta \mathcal{Q}(\theta \mid y) \pi_1(\theta) d\theta} \right) \frac{p_0}{p_1} \qquad (11)$$
$$= B(y)\Omega$$

where $L(\theta|y) \propto Q(\theta|y)$ has been used to eliminate the explicit normalizing constant. The last two lines of Eq. 11 define the Bayesian odds adjustment factor, or odds ratio, $B(y)$. Note that, unlike $\mathcal{Q}(\theta|y)$, $B(y)$ is not completely objective, since prior probability distributions are required to calculate it. Applications of this formula are referred to as Bayesian hypothesis testing, as distinct from the Bayesian parameter evaluation described in previous sections.

In the general context of Bayesian hypothesis testing there can arise an oddity in the statistical inference between the two alternatives. When a point or very narrow null hypothesis π_0 is being tested against a diffuse or vaguely characterized alternative hypothesis π_1 Bayesian hypothesis testing may lead to an unreasonable result in which data internally quite distinct from the null hypothesis are nevertheless regarded as supporting the null in preference to the alternate. Mathematically, a likelihood $\mathcal{Q}$ whose maximum is several standard deviations away from the null still yields $B(y) > 1$. This situation is referred to by various authors as *Jeffreys' paradox* or Lindley's paradox. It is well described by G. Shafer (1982):

"Lindley's paradox is evidently of great generality; the effect it exhibits can arise whenever the prior density under an alternative hypothesis is very diffise relative to the power of discrimination of the observations. The effect can be thought of as an example of conflicting evidence: the statistical evidence points strongly to a certain relatively small set of parameter values, but the diffise prior density proclaims great skepticism (presumably based on prior evidence) towards this set of parameter values. If the prior density is sufficiently diffuse, then this skepticism will overwhelm the contrary evidence of the observations.

"The paradoxical aspect of the matter is that the diffise density $\pi_1(\theta)$ seems to be skeptical about all small sets of parameter values. Because of this, we are somewhat uneasy when its skepticism about values near the 'observed interval' overwhelms the more straightforward statistical evidence in favor of those values. We are especially uneasy if the diffuseness of $\pi_1(\theta)$ represents weak evidence, approximating total ignorance; the more ignorant we are the more diffise $\pi_1(\theta)$ is, yet this increasing diffuseness is being interpreted as increasingly strong evidence against the 'observed interval.' "

Shafer's article then proceeds to a cogent argument that cases where a Lindley paradox occurs are precisely those where ordinary Bayesian hypoth-

esis testing is misleading and should not be used. (In fairness, one should note that the major development of Shafer's treatment is an extension of Bayesian formalism to deal with this awkward case; and that the published article includes an assortment of counter-arguments from various authors.) The problem, of course, is that a diffuse prior is being treated as evidence against the hypothesis in question. As noted above, $B(y)$ is not an objective adjustment of subjective prior odds between hypotheses, but depends on a second subjective choice of prior distribution for an alternate. (If the null hypothesis is also not well defined, it presents yet a third opportunity for subjective contributions.) If not carefully noted, these further subjective elements can be quite as inexplicit and misleading as those that Bayesians object to in classical analvsis.

A further practical difficulty with hypothesis testing, relative to classical treatments, is that the null hypothesis is always compared to a *specific* alternative. In many situations, including the PEAR experiments, investigators are interested in any possible deviation from a specified range of possibilities, without having enough information about the possible character of such a deviation to construct one specific alternate with any degree of conviction. A diffuse alternative that encompasses a wide range of probabilities is not a satisfactory option. This can be seen abstractly, from consideration of the Lindley paradox in cases where the statistical resolving power of a proposed experiment will be very high; it can also be argued on other grounds, as will be discussed below under the heading of statistical power.

Hypothesis Testing on PEAR Data

The extreme sharpness of the likelihood function for the PEAR data base used earlier makes any hypothesis test on it susceptible to a Lindley paradox. Unless the prior for the alternate is also narrowly focused in the region of high likelihood, $B(y)$ will claim unreasonable support for the null hypothesis.

One might then argue that the recipe for avoiding dubious results is to employ a narrow range of values for the alternate hypothesis. There are, after all, numerous arguments that anomalous effects such as PEAR examines should be small. Perhaps the simplest argument is that if such effects were large, they would not be a subject of dispute! Despite such reasoning, as recently as 1990 an article appeared using, at one point, $\pi_1(p) = 1, p \in [0, 1]$ for the alternate in a hypothesis test (Jeffreys 1990). The use of highly diffuse priors can thus be seen to be a real and current practice meriting cautious examination, rather than a purely argumentative point.

The final section of the parameter-evaluation discussion above, with its two-component prior, is already very close to a hypothesis test. The only major difference is that the weighting parameter a is absorbed into the odds Ω, leaving a one-parameter family of alternate priors for comparison with the null. Figure 6 shows the value of B for a range w values (where w is the width of the rectangular alternate). The solid line, marked "Symmetric", can be

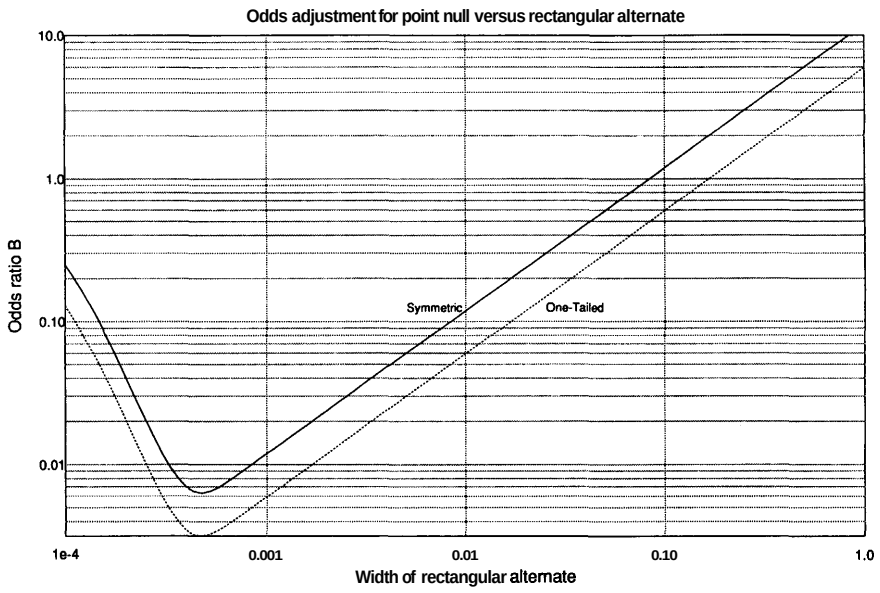

Fig. 6.

seen to be the limit of the w-dependence shown in Figure 5 for $a \rightarrow 0$. The dotted line, marked "One-Tailed", shows the odds ratio for the null against a one-sided version of the rectangular prior, which has support only for $p > 0.5$. Since the PEAR results are based on a directed hypothesis, one-tailed statistics are appropriate in a classical framework, and this would seem to be an appropriate Bayesian analog, as well. Both functions attain a minimum at $w \approx 4.8 \times 10^{-4}$, for $B = 0.00316$ in the one-tailed case.

Inflation of p-values and Statistical Power

The smallest B factor in the hypothesis comparison above was a factor of 10 larger than the two-tailed p-value of 3×10^{-4} quoted in *Margins.* The smallest *B* to emerge from a direct hypothesis test for these data is 0.00146, for comparison of a point null against a point alternate located exactly at the maximum likelihood $p = s/n$. This is still a factor of 10 larger than the corresponding one-tailed value (the Bayesian test is also "one-tailed" in this case). The tendency of hypothesis comparison to emerge with a larger B value than the corresponding p-value of a classical test is often cited by Bayesian analysts as evidence that classical p values are misleading for large databases, and should be adjusted by some correction factor, perhaps of order $n^{1/2}$. (See, for example, the discussions by Good and Hill in the latter portions of the Shafer article; see also Jeffreys (1990).) Such proposals generally fail to take

into account considerations of statistical power, a somewhat neglected branch of analysis.

Conventional statistical reasoning recognizes two types of errors. The more commonly acknowledged Type I or a error is the false rejection of the null hypothesis, where a is the probability of making such an error. Type **II** or β error is the false ***acceptance*** of the null hypothesis, with β likewise being the probability of making the error. $1 - \beta$ is usually called the statistical power of a test. In any real situation, the null hypothesis is either true or false and therefore only one of the two types of error is actually possible. **A** less obvious point is made in the literature:

"The null hypothesis, of course, is usually adopted only for statistical purposes by researchers who believe it to be false and expect to reject it. We therefore often have the curious situation of researchers who assume that the probability of error that applies to their research is β (that is, they assume the null hypothesis is false), yet permit β to be so high that they have more chance of being wrong than right when they interpret the statistical significance of their results. While such behavior is not altogether rational, it is perhaps understandable given the minuscule emphasis placed on Type II error and statistical power in the teaching and practice of statistical analysis and design . . ." (Lindsey 1990)

Consider an experiment involving N Bernoulli trials where one wishes to know whether they are evenly balanced ($p = 0.5$, the null hypothesis) or biased, even by extremely small deviations from the null hypothesis. (This is in fact the case in PEAR REG experiments.) Consider two cases: the null hypothesis is true ($p = 0.5000$); the null hypothesis is false with $p = 0.5002$. Assume that the experiment (in each case) is analyzed by two statisticians, neither of whom has any advance knowledge of ***p:*** a classical statistician who rejects the null hypothesis if a two-tailed p-value ≤ 0.01 is attained, and a Bayesian who, using a uniform prior for the alternate, regards the experiment as supporting the null hypothesis if the odds adjustment factor $B > 1$, and as supporting the alternate if $B < 1$. To give the probability estimates some concrete reality we may imagine the experiment being run many times with different pairs of analysts. The probability that the classical statistician makes a type I error is defined by the choice of a, and is independent of N. The table below gives the probability, for various N, of a type I error by the Bayesian analyst (regarding the evidence as favoring the alternate when the null is true) and the probability of type II error by either analyst. For either a true null or a true alternate, the final experimental scores follow a binomial distribution with N determined by the row of the table and p = .5000 or .5002 respectively. For both the classical analyst and the Bayesian analyst, one may calculate the number of successes needed for an analyst to reject the null hypothesis g, where the Bayesian is regarded as rejecting the null if $B < 1$. The table then quotes the error frequencies that follow from the actual sucess distributions under each hypothesis and the analytical criterion used for rejection. The

TABLE I
Error rates under different analyses

N	Null is true	Null is false	Null is false
	a error, Bayesian	**β error, classical**	**β error, Bayesian**
100	0.028	0.995	0.982
10,000	0.0031	0.994	0.998
10^6	2.6×10^{-4}	0.985	0.999
10^7	7.6×10^{-5}	0.905	0.906
10^8	2.2×10^{-5}	0.077	0.595
1.5×10^8	1.86×10^{-5}	0.010	0.267
10^9	6.8×10^{-6}	3.6×10^{-24}	1.8×10^{-16}

probabilities of type II error combine the probability of erroneous acceptance of the null hypothesis with that of (correct) rejection of the null due to mistakenly inferring $p < 1/2$. For the considerations of columns 3 and 4, $p > 1/2$, and both conclusions are equally erroneous. The abrupt drop of β values in the last few lines of the table may seem jarring, but is a rather generic feature of power analysis. For any given constant effect size, there will be a fairly narrow range of N (as measured on a logarithmic scale) for which any specific test will quickly shift from being almost useless to being virtually certain to spot the effect.

A salient feature is that the Bayesian calculation, with this prior, starts out more vulnerable to type I error, and less vulnerable to type II error, for small N: however, they are both so likely to suffer type II error that this is not very interesting. For large N, the Bayesian calculation is uniformly more conservative in that its probability of falsely rejecting the null hypothesis declines with N, while the classical analysis uses a constant p-value criterion for rejecting the null. Correspondingly, the Bayesian calculation has a far higher likelihood than the classical of falsely accepting the null hypothesis. The row for $N = 1.5 \times 10^8$ is of special interest, because for this value the classical analysis attains equal likelihood of type I and type II errors. At this level the Bayesian analysis still has over 1 chance in 4 of incorrectly confirming $p = 0.5$.

Table I actually makes an extremely generous interpretation of the Bayesian output. The Bayesian analyst is assumed to regard the data as supporting the alternate hypothesis as soon as the Bayes factor $B < 1$. However, as mentioned above, various authors write as though the odds adjustment factor B ought to be regarded as an analogue to the p-value for a data set. Had this sort of reasoning been used in constructing Table I, the Bayesian analyst would still have $p = 0.634$ for committing a type II error on 150 million trials.

The Bayesian analysis used is not optimized for the problem of testing, say, a circuit that produces "on" signals with a probability that is definitely either $p = 0.5$ or $p = 0.5002$. Neither is the classical analysis. If the problem were to distinguish these two discrete alternatives, a Bayesian test would compare two point hypotheses; while a classical test might, with given reliability levels,

establish ranges of output for which a circuit would be classed as "definitely 0.5", "definitely 0.5002", or "inconclusive, further testing required." The actual problem may be envisioned as a sociological thought experiment in which large numbers of Bayesian and classical analysts are presented only with the output of the device, and the information that the underlying Bernoulli probability either was or was not 0.5. The uniform Bayesian alternate simply represents ignorance of possible alternative values of p , and is directly analogous to the situation, described earlier, for selection of priors in anomalies data. The second to last line of Table I says that, were such an experiment conducted and each analyst presented with 150 million trials with either $p = 0.5$ or $p = 0.5002$, the classical analysts would produce 1% false positives and 1% false negatives; while the Bayesian analysts would produce a vanishingly tiny fraction of false positive reports but over 26% false negatives—deviant datasets identified as unbiased.

In a more general vein, for large databases with small effects, it is apparent in light of the various discussions above that *any* Bayesian hypothesis comparison will yield an odds adjustment factor larger than the classical p-value for the same data. If the odds adjustment B is regarded as equivalent to a p-value, or a corrected version of it, the inevitable consequence will be a test less powerful than the classical version, and so more prone to missing actual effects that may be present for any given database size.

An important consideration in statistical power analysis is the effect size. One seldom has the advantage of knowing in advance the magnitude of potential effects. In the anomalies research program at PEAR, for example, *any* unambiguous deviation from the demonstrable null hypothesis range has profound theoretical and philosophical import. While traditional power analysis would suggest scaling the sample sizes to the smallest effect clearly distinguishable from the null hypothesis range, this would be totally impractical in that it would require datasets several orders of magnitude larger than those published in *Margins.* This, too, is a standard situation frequently encountered in power analysis, in that effects of potential interest may nonetheless be too small to identify in studies of manageable size. While in fact the apparent effect size manifest in the PEAR data is much larger than this pessimistic case, there was no way of knowing in advance that this would be so. Confronted with the possibility of very small effects, the only viable alternative may be to conduct such measurements as are feasible, with the awareness that effects may be too small to measure in the current study; in which case the experiment will at least permit the establishment of an upper bound to the effect in question.

In such a situation, a Bayesian analysis using the uniform alternate prior is obviously too obtuse to be of value; it retains a high chance of a false negative report for dataset sizes where the classical test has a high degree of reliability. At the same time, information about plausible alternates may well be so scant that the uniform prior, or an only slightly narrower one, is nonetheless a fair summary of one's prior state of knowledge. Under these circum-

stances, the reasonable course would seem to be adoption of classical statistical tests with an experiment designed to exclude any procedures, such as optional stopping, which would invalidate the tests. The next section will discuss optional stopping and related issues in more detail.

Relative Merits: Bayesian vs. Classical Statistics

Bayesian analysis is occasionally claimed to remedy various shortfalls in the classical analysis of very large data bases (Jeffreys, 1990; see also Utts, 1988). Beyond the question of replacing classical p values with Bayesian odds adjustment factors discussed above, two other sources of inadequacy are usually cited: First, any repeated measurement eventually reaches a point of diminishing returns where further samples only refine measurement of systematic biases rather than of the phenomenon under investigation. Second, indefinite continuation of data collection guarantees that arbitrarily large excursions will eventually arise from statistical fluctuations ("sampling to a foregone conclusion"). Both of these concerns, together with the notion that Bayesian analysis is specially qualified to deal with them in a way that classical analysis is not, are not substantiated by well-designed REG experiments in general, or by the *Margins* data in particular.

1. *The inevitable dominance of bias.* The maximum possible influence of biasing effects in this experiment has been discussed in the context of the "null hypothesis interval" above, and displayed graphically in Fig 1. In an experiment that contrasts conditions where the only salient distinction is the operator's stated intention, any systematic technical error must itself correlate with intention to affect the final results. While unforseen effects may never be completely ruled out, it would require considerable ingenuity to devise an error mechanism that achieved this correlation without itself being as anomalous as the putative effect. Over the eight years of experimentation that went into the *Margins* database (twelve years as of this writing), both the PEAR staff and interested outsiders, including prominent members of the critical community, have been unable to find any such mundane source of systematic error. Beyond this, the bias question in REG data is an improper conflation of two unrelated issues. As pointed out by Hansel (1966) in the evaluation of any data, a statistical figure-of-improbability measures only the likelihood that data are the result of *random* fluctuation. It remains for each analyst to draw conclusions as to whether the deviation from expected behavior is more plausibly due to the effect under investigation or to an unaccounted-for systematic bias in the experiment. Thus, the question of bias is essentially external to the purely statistical issue of whether or not the data, are consistent with a null hypothesis.
2. *Arbitrarily large excursions.* The conclusion of Feller's (1957) discussion of the law of the iterated logarithm may be summarized thus: Any

threshold condition for the terminal Z score of a binary random walk that grows more slowly than $\sqrt{2log(log(n))}$ will be exceeded infinitely many times as the walk is indefinitely prolonged, and thus is guaranteed to be exceeded for arbitrarily large data bases. Obviously, this is of concern only for experimental sequences of indeterminate length, where one could, in principle, wait for one of these large excursions to occur, and then declare an end to data collection. Any experiment of predefined length will always have a well-defined terminal probability distribution. Without exception, all PEAR laboratory data, including the Margins array, have conformed to the latter, specified series length protocols. Nevertheless, if the Margins data are arbitrarily subjected to a worst-case, "optional-stopping-after-any-trial" analysis, the probability that a terminal Z score of 3.014 could be attained at any time in the program's history computes to ≤ 0.007. Under the somewhat more realistic assumption that data accumulation could be halted only after any of the 87 series that comprise the database, the terminal probability becomes 50.002. The actual history of the experimental program clearly demonstrates that no optional stopping strategy can have been applied to publication decisions, for significant effects have steadily been apparent from the collection of the first series onward (Dunne, Jahn, and Nelson, 1981), and the various publication points have never coincided with local maxima in the accumulating results (Jahn 1982; Dunne, Jahn, and Nelson 1982; Jahn, Dunne, and Nelson 1983; Nelson, Dunne, and Jahn 1984; Nelson, Jahn, and Dunne 1986; Jahn and Dunne 1986; Jahn 1987). For classical tail-measurement statistics to be legitimate, it is sufficient that the termination condition be independent of the outcome of the experiment (Good 1982).

3. Special competence of Bayesian analysis. The likelihood function of Bayesian analysis will as a rule replicate the results of a classical analysis in the sense that, if classical statistics compute a Z score z, then the likelihood function will have a mean that is z of its own standard deviations away from a point null hypothesis. (This follows from the functional form of Q.) Any differences in interpretation must therefore come from the use of different priors. We have seen above that Bayesian parameter evaluation with a prior that is uniform in the region of high likelihood likewise replicates the classical analysis, since this creates a posterior probability with the same shape as the likelihood. While non-uniform priors can change this conclusion, for a likelihood function as sharply focused as that of the Margins data such priors must be close to the null-hypothesis delta function, i.e., recalcitrant almost to the point of impenetrability, before they force acceptance of the null hypothesis. Direct comparison of competing hypotheses, on the other hand, is vulnerable to confounds from inappropriate alternate priors.

As already noted, no analysis, Bayesian or otherwise, will guard indefinitely against an unforseen bias. Table I and the related discussion showed that

Bayesian analysis with a recalcitrant prior eventually agrees with classical analysis in rejecting the null hypothesis, when enough data are accumulated with a constant mean shift. They also show an example, appropriate to the class of REG-type experiments, where Bayesian analyses that choose priors to be very conservative are also necessarily very insensitive and must suffer a large probability of type II error. This is true whether the effect is real or a systematic error, and the mode of analysis grants no special ability to distinguish the two cases.

Data Scaling

A final point of comparison concerns the interpretation of the *Margins* data on various scales. Classical analysis does not require that any special attention be paid to the intermediate structure of the experimental data; if a Z score is computed for each series, and the assorted series Z scores are compounded appropriately, the composite result is exactly the Z that would result were the data treated in aggregate. This occurs because, no matter what scale is used to define elements of the data, the increased consistency of the results exactly compensates for the loss of statistical leverage from the decreased N of units. Processing REG data in large blocks is essentially a signal averaging procedure, unusual only in that it is performed algorithmically on stored data rather than in preprocessing instrumentation.

Directly checking for the same sensible scaling property in Bayesian analysis would entail developing an extension of the formalism for continuously distributed variables, beyond the scope of this discussion. However, a cursory look at the issue can be accomplished by examining the series data breakdown in *Margins*. The column listing $p < 0.5$ allows the 87 series to be regarded as 87 Bernoulli trials, each one returning a binary answer to the question, "Did the operator achieve in the direction of intent, or not?" Naturally a great deal of information is lost in this representation, since the differential degree of success or failure cannot be reckoned, but it remains instructive. Of the 87 series, 56 were "successes" as Bernoulli trials. The binomial distribution for 87 p = 0.5 trials has $\mu = 43.5$, a = 4.66. The actual success rate thus translates to z = 2.68, p = 0.004 one-tailed. The loss of information is seen in the reduction of significance, but the result is consistent in being a strong rejection of the null.

Not so for a Bayesian hypothesis test against the uninformed alternate $\pi_1 = 1$. For the binary data, as we saw, B = 12; but $B(87, 56) = 0.2$. Where the reduced information decreased the significance of the classical result, as one might expect, it has inverted the Bayesian result from a modest confirmation of the null to a modest rejection of the null. The discrepancy, of course, lies in the Lindley paradox: the naive alternate prior is inappropriate for the binary test, but not unreasonable for the vastly larger effect *that must be present, if the effect is real* on the series scale. The fact that the inversion occurs is itself confirmatory evidence for the reality of the mean shift and therefore evidence

against the utility of a test that regards the data as supporting the null hypothesis.

Final Comments and Summary

The main points to emerge from this study are:

1. For a Bayesian analysis of Bernoulli trials an objective likelihood function can be constructed which obeys the necessary addition rule for consistent handling of accumulating data (Eq 6). The likelihood function has the same distribution as a classical estimate of confidence intervals on the value of p , and differences of interpretation can therefore arise only from the choice of priors.
2. It therefore follows that a prior that is uniform in the region of high likelihood, thus producing a posterior probability of the same shape as the likelihood, replicates the classical analysis. For the PEAR data, this reproduces a two-tailed p of 3.0×10^{-4} against a point null hypothesis.
3. Prior belief favoring the null hypothesis impacts the conclusions. In its ultimate expression, where only values consistent with the null hypothesis are allowed prior support, no evidence can sway the outcome. Less extreme forms continue to reject the null hypothesis (exclude it from reasonable parameter confidence intervals) unless the prior includes much of the probability within the null (thus approaching the impervious case) or is narrower than the likelihood (and therefore narrower than the statistical leverage of the known number of trials justifies.) Some examples using the PEAR database include: Exclusion of the null hypothesis from at least the 95% posterior confidence interval for a normal prior centered on the null hypothesis with a as small as 3×10^{-5}, compared to $a = 4.9 \times 10^{-5}$ for the likelihood function; posterior odds against the null hypothesis of 20 to 1 for a prior that starts with 85% of the strength concentrated at $p = 0.5$ and the remainder uniformly distributed with width 5×10^{-4}.
4. Hypothesis comparison needs to be approached with caution because the odds adjustment factor $B(y)$ contains a contribution from the choice of prior probability distributions, and so is at least as vulnerable to prior prejudices as are the prior odds Ω (Eq. 9)
5. Hypothesis tests return odds correction factors larger than classical p-values even when near-optimal cases are chosen. For the PEAR data, the optimal case is a comparison of a point null against a point alternate at $p = 0.50018$ (the maximum of the likelihood function), leading to $B = 0.00146$ (odds of 685 to 1 against the null.) A consideration of statistical power, however, demonstrates that this does not establish a flaw in, or correction to, classical p-values but is a simple consequence of adopting a less sensitive test.
6. Examination of the response of Bayesian hypothesis testing to large

databases indicates that claims of special ability to deal with biases or optional stopping, or of qualitatively superior response to increasing amounts of data, compared to classical statistics, are unwarranted.

7. In a situation such as confronted by the PEAR program and related investigations, where any detectable effect is of fundamental interest and importance, the necessity of having a specific alternate hypothesis for a Bayesian hypothesis test is a limiting and potentially confounding factor. A prior that is diffuse enough to reflect ignorance of potential effects will have much less statistical power than an appropriate classical test.

Thus, while Bayesian statistical approaches have the virtue of making their practitioner's prejudices explicit, they may in some applications allow those prejudices more free rein than is usually acknowledged or desirable. Whereas a classical analysis returns results that depend only on the experimental design, Bayesian results range from confirmation of the classical analysis to complete refutation, depending on the choice of prior. Those priors that disagree strongly with the classical analysis frequently show one or more suspect features, such as being either very diffuse or pathologically concentrated with respect to the likelihood. (While it violates the definition of a prior to adjust it with respect to an observed effect, the *width* of the likelihood is determined only by the experiment's size, not its outcome, and is therefore a legitimate guide to the characteristics of reasonable priors.) This would suggest that, the more strongly a Bayesian analysis disagrees with a classical result, the more likely the disagreement is due to a subjective contribution of the analyst.

Given the impact of prior probabilities, one might argue that the proper role of a Bayesian analysis should be strictly to quote likelihood functions and allow each reader to impose his own priors. However, the philosophical exercise of justifying (or refuting) various priors remains a valuable one, particularly for clarifying the meaning of a particular result.

Acknowledgements

The author wishes to thank Robert Jahn, Brenda Dunne, Roger Nelson and Elissa Hoeger for reading earlier drafts of the manuscript and providing valuable advice. The Engineering Anomalies Research program is supported by major grants from the McDonnell Foundation, the Fetzer Institute, and Laurance S. Rockefeller.

Correspondence and requests for reprints should be addressed to York Dobyns, Princeton Engineering Anomalies Research, C-131 Engineering Quadrangle, Princeton University, Princeton, New Jersey 08544-5263.

References

Dunne, B. J., Jahn, R. G., & Nelson, R. D. (1981). *An* REG *Experiment with Large Data-Base Capability.* Princeton University: PEAR Technical Note.

Dunne, B. J., Jahn, R. G., & Nelson, R. D. (1982). *An* REG *Experiment with Large Data Base*

Capability, ZZ: Effects of Sample Size and Various Operators. Princeton University: PEAR Technical Note.

Feller, W. (1957). *An Introduction to Probability Theory and Its Applications.* Volume 2. (2nd ed.) New York, London: John Wiley & Sons.

Good, I. J. (1982). Comment [on Shafer (1982), see below]. *Journal of the American Statistical Association,* 77, 342.

Hansel, C. E. M. (1966). *ESP / A Scientific Evaluation.* New York: Charles Scribner's Sons.

Jahn, R. G. (1982). The persistent paradox of psychic phenomena: An engineering perspective. *Proceedings of the ZEEE,* 70, 136–170.

Jahn, R. G. (1987). Psychic Phenomena. In G. Adelman, Ed., *Encyclopedia of Neuroscience, Vol. II.* pp. 993–996. Boston, Basel, Stuttgart: Birkhauser.

Jahn, R. G., & Dunne, B. J. (1986). On the quantum mechanics of consciousness, with application to anomalous phenomena. Foundations of Physics, 16(8) 721–772.

Jahn, R. G., & Dunne, B. J. (1987). *Margins of Reality.* San Diego, New York, London: Harcourt Brace Jovanovich.

Jahn, R. G., Dunne, B. J., & Nelson, R. D. (1983). Princeton Engineering Anomalies Research. In C. B. Scott Jones, Ed., *Proceedings of a Symposium on Applications of Anomalous Phenomena, Leesburg, VA., November 30-December 1, 1983.* Alexandria, Santa Barbara: Kaman Tempo, A Division of Kaman Sciences Corporation.

Jahn, R. G., Dunne, B. J., & Nelson, R. D. (1987). Engineering Anomalies Research. *Journal of Scientific Exploration,* 1(1), 21–50.

Jeffreys, W. H. (1990). Bayesian Analysis of Random Event Generator Data. *Journal of Scientific Exploration,* 4(2), 153–169.

Lipsey, M. W. (1990). *Design Sensitivity: Statistical Power for Experimental Research.* Newbury Park, London, New Delhi: SAGE Publications.

Nelson, R. D., Bradish, G. J., & Dobyns, Y. H. (1989). *Random Event Generator Qualification, Calibration, and Analysis.* Princeton University: PEAR Technical Note.

Nelson, R. D., Dunne, B. J., & Jahn, R. G. (1984). *An* REG *Experiment With Large Data Base Capability, ZIZ: Operator Related Anomalies.* Princeton University: PEAR Technical Note.

Nelson, R. D., Jahn, R. G., & Dunne, B. J. (1986). Operator-related anomalies in physical systems and information processes. *Journal of the Society for Psychical Research.* S3(803) 261–286.

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

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

Journal of Scientific Exploration, Vol. 6, No. 1, pp. 47–57, 1992 0892-3310/92

Response to Dobyns

WILLIAM H. JEFFERYS

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

Abstract– 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 $p, = 0.5$ is 1.1×10^{-6}. In using a uniform prior, with $w = 10^{-3}$ (as suggested by 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 $p, = 0.5$ can adequately represent the views of a person who has 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 considerationslead to a Bayesian justification of Ockham's razor. See Jaynes (1979), Smith and Spiegelhalter (1980), Gull (1988), Loredo (1990), Berger and Jefferys (1992), Jefferys and Berger (1992), 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 $\mathbf{0} = |p - p_{\Delta}|/p_{\Delta}$ is very small.

Whether the discrepancyis 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 accommodate 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-Pearsonerror probability,a, 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 significancelevel. 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. A 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_1, and reject the alternative if its posterior probability becomes less than p,) (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. 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, shufflethem 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 detectability, 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. (1992). "Minimal Bayesian testing of precise hypotheses, model selection, and Ockham's razor." To appear in *Journal of the Italian Statistical Society*.

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. (1992) "Ockham's Razor and Bayesian Analysis." *American Scientist*, 80, 64–72.

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.