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 θ. One's state of prior knowledge, or ignorance, may be expressed as a prior probability distribution $\pi_0(\theta)$ over the space of possible θ 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, θ 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

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

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, $\mathcal{L}$. The general recipe for updating one's knowledge of parameter(s) θ in light of evidence y is expressed by

$$\pi_1(\theta|y) \propto \mathcal{L}(\theta|y)\pi_0(\theta) \tag{2}$$

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

$$\pi_1(p\,|\,n,\,s) \propto \mathfrak{L}(p\,|\,n,\,s)\pi_0(p) = \binom{n}{s}p^s(1-p)^{n-s}\pi_0(p). \tag{3}$$

Note that (2) and (3) are expressed as proportionalities rather than equalities. While there is always a normalization for $\mathfrak{L}$ such that $\mathfrak{L}(\theta\,|\,y)\pi_0(\theta)$ has a total integral of 1, this normalization in general depends on $\pi_0(\theta)$. Specifically, the use of $L(\theta\,|\,y) = \mathfrak{L}(\theta\,|\,y)/\int_\theta d\theta \mathfrak{L}(\theta\,|\,y)\pi_0(\theta)$, where $\int_\theta$ denotes integration over all possible values of θ, produces the correct normalization. The relations (2) and (3), on the other hand, express $\mathfrak{L}$ purely as a function of y and θ without reference to prior probabilities. The conceptual clarity of stepping from one state of knowledge about θ, 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

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

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,

$$\pi_2(\theta\,|\,y_1,\,y_2) \propto \mathfrak{L}(\theta\,|\,y_2)\pi_1(\theta\,|\,y_1) \propto \mathfrak{L}(\theta\,|\,y_2)\mathfrak{L}(\theta\,|\,y_1)\pi_0(\theta). \tag{5}$$

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 $\pi_2(\theta\,|\,y_1,\,y_2) \propto \mathfrak{L}(\theta\,|\,y_1 + y_2)\pi_0(\theta)$. 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 $\mathfrak{L}$ has the property $\mathfrak{L}(\theta\,|\,y_1 + y_2) \propto \mathfrak{L}(\theta\,|\,y_1)\mathfrak{L}(\theta\,|\,y_2)$. The likelihood function $\mathfrak{L} = p^s(1-p)^{n-s}$ indeed has this essential addition property:

$$\mathfrak{L}(p\,|\,n_1 + n_2,\,s_1 + s_2) = p^{s_1+s_2}(1-p)^{(n_1+n_2)-(s_1+s_2)} = \mathfrak{L}(p\,|\,n_2,\,s_2)\mathfrak{L}(p\,|\,n_1,\,s_1). \tag{6}$$

Since Bayesian formalism is internally consistent, a calculation such as (6) is essentially a check on the validity of the likelihood function; any legitimate $\mathfrak{L}$ 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 "ΔPK" 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 $\mathcal{L}(p|n, s) = p^s(1 - p)^{n-s}$. The mean of this p distribution is $(s + 1)/(n + 2)$; for large n its standard deviation is a $\sigma = \dfrac{1}{2\sqrt{n}} + O(1/n)$; and it becomes essentially normal. With the values above,

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

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 possibilities are 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 + δ, 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 "ΔPK" 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 ΔPK results, any residual artifactual bias of the device or the data processing would itself have to correlate with the operator's intention. Specifically, if $p_0 = 0.5 + \delta$ 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 ΔPK of:

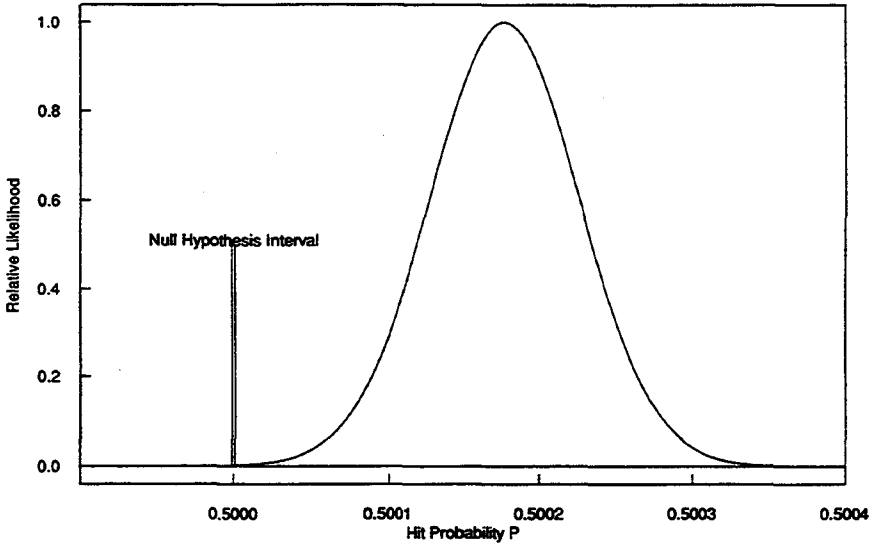Fig. 1. Likelihood function for PEAR REG data.

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

Thus, when $N_h = N_l$, $p_\Delta \equiv 0.5$, regardless of the value of δ. 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 $|\delta| < 0.0002$ 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:

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

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
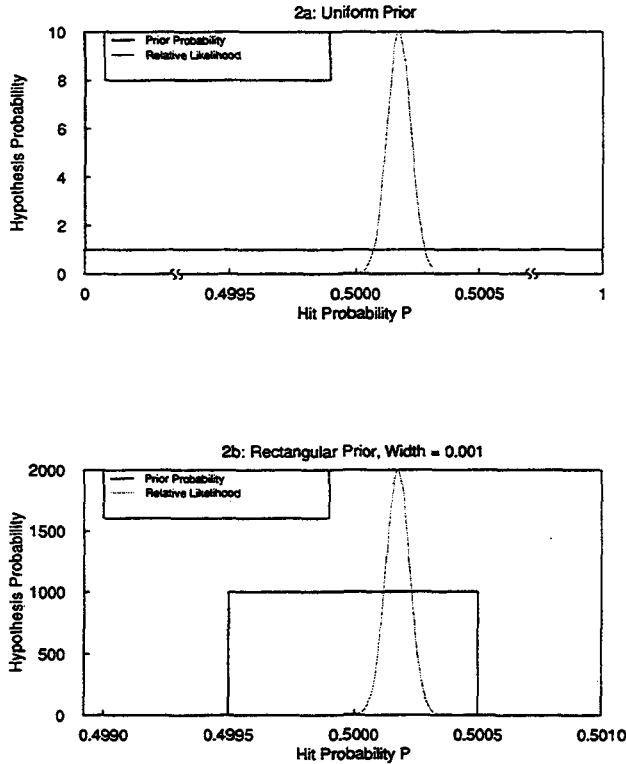
2a: Uniform Prior

2b: Rectangular Prior, Width = 0.001

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 considering another family of priors that specifically favor the null hypothesis to varying degrees but do not have sharp cutoffs of probability. Let $\pi_0(k, p) = [(2k + 1)!/(k!)^2]p^k(1 - p)^k$ for any k. All of these functions are properly normalized probability distributions, with mean 0.5 and standard deviation $\sigma = \frac{1}{2}\sqrt{(k - 1)/(2k^2 + 5k + 3)}$, 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
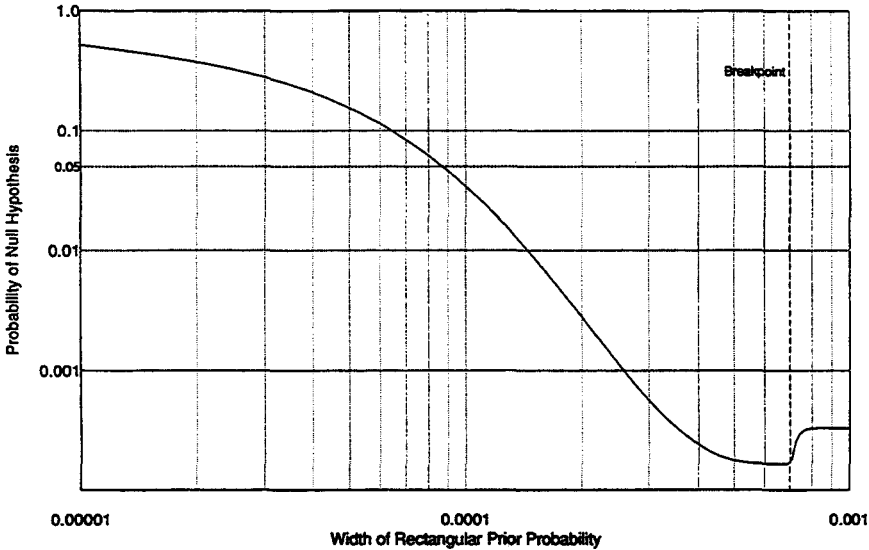
Fig. 3.   Conjugate confidence intervals from rectangular priors.

the *Margins* data, the resulting π_1 has mean $\mu_1 = (s + 1 + k)/(n + 2 + 2k)$ and standard deviation $\alpha_1 = \frac{1}{2}\sqrt{n + 2k}$ (in the large-n approximation, which is clearly justified), as can be seen from the functional form of ℓ and the fact that multiplying by π_0 is equivalent to the substitution $s \rightarrow s + k$, $n \rightarrow n + 2k$, 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 $k \rightarrow \infty$, 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 $\sigma < 3 \times 10^{-5}$, is required. Here the characteristic scale 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
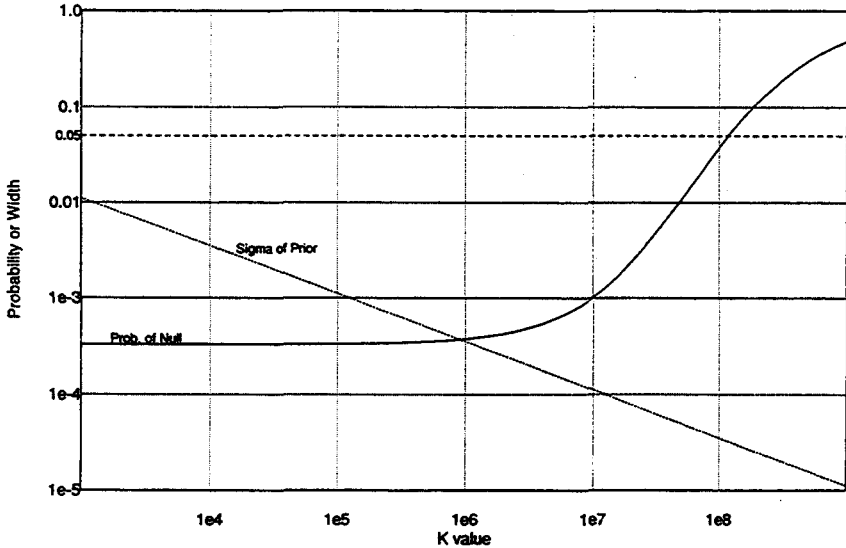
Fig. 4. Conjugate probabilities from K-family of priors.

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 $\pi_0(a, w, p) = a\delta(p - 0.5) + (1 - a)\pi_w(p)$, 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

Posterior Probability of Null Component

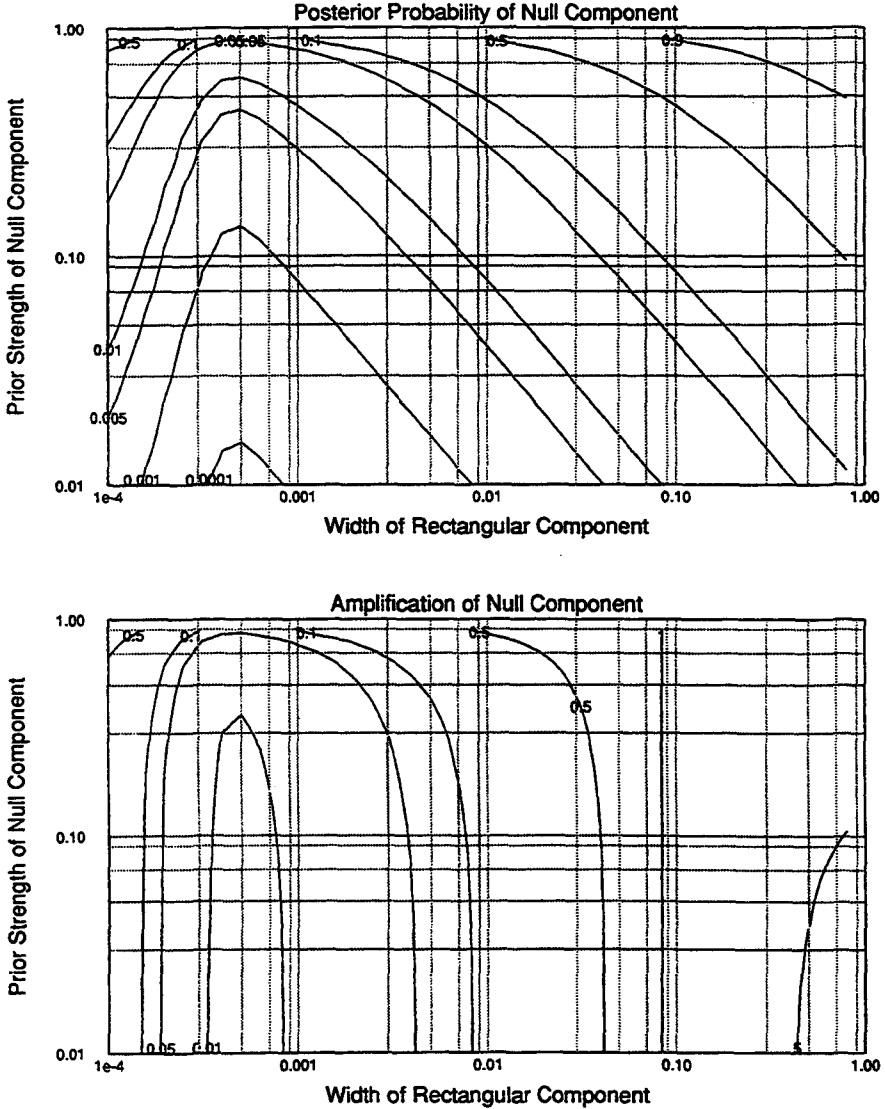Amplification of Null Component

Fig. 5.

amplification factor tends toward a constant depending only on w, and that even for $a \approx 0.85$, 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

$$\Omega'(\theta \mid y) = \frac{p_0'}{p_1'} = \frac{\displaystyle\int_\theta L(\theta \mid y)p_0\pi_0(\theta)d\theta}{\displaystyle\int_\theta L(\theta \mid y)p_1\pi_1(\theta)d\theta}$$

$$= \left(\frac{\displaystyle\int_\theta \ell(\theta \mid y)\pi_0(\theta)d\theta}{\displaystyle\int_\theta \ell(\theta \mid y)\pi_1(\theta)d\theta}\right)\frac{p_0}{p_1} \tag{11}$$

$$= B(y)\Omega$$

where $L(\theta \mid y) \propto \ell(\theta \mid y)$ 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 $\ell(\theta \mid y)$, $B(y)$ 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 ℓ 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 diffuse 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 diffuse 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 diffuse 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 diffuse $\pi_1(\theta)$ 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 analysis.

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
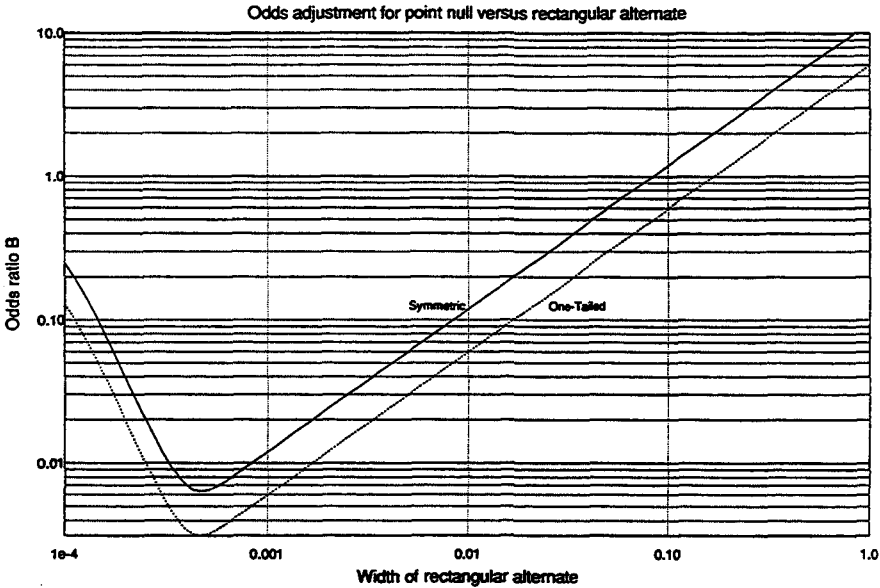
Fig. 6.

seen to be the limit of the w-dependence shown in Figure 5 for $a \to 0$. 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 α error is the false rejection of the null hypothesis, where α 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 α, 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
	α 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 mis-
takenly 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 conser-
vative 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 ≤ 0.002. 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 $\mathcal{L}$.) 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 $\mu = 43.5$, $\sigma = 4.66$. 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 σ as small as 3×10^{-5}, compared to $\sigma = 4.9 \times 10^{-5}$ 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.

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, II: 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 IEEE,* 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, III: 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.