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

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

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

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

This is graphed in Figure 1.

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

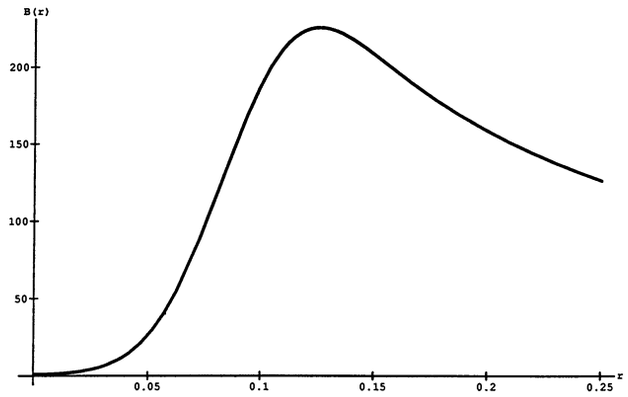

FIG. 1. *The Bayes factor of H_1 to H_0 as a function of r, the maximum change in success probability that is expected given that ESP exists, for the ganzfeld experiment.*

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

ACKNOWLEDGMENTS

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

Comment

Ree Dawson

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

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

bate. This debate is also a good example of how statistical criticism can be part of the scientific process and lead to better experiments and, in general, better science.

The remainder of the paper addresses technical issues of meta-analysis, drawing upon recent research in parapsychology for an in-depth application. Through a series of examples, the author presents a convincing argument that power issues cannot be overlooked in successive replications and that comparison of effect sizes provides a richer alternative to the dichotomous measure inherent in the use of p-values. This is particularly relevant when the potential effect size is small and resources are limited, as seems to be the case for psi studies.

The concluding section briefly mentions Bayesian techniques. As noted by the author, Bayes (or empirical Bayes) methodology seems to make sense for research in parapsychology. This discussion examines possible Bayesian approaches to meta-analysis in this field.

BAYES MODELS FOR PARAPSYCHOLOGY

The notion of repeatability maps well into the Bayesian set-up in which experiments, viewed as a random sample from some superpopulation of experiments, are assumed to be exchangeable. When subjects can also be viewed as an approximately random sample from some population, it is appropriate to pool them across experiments. Otherwise, analyses that partially pool information according to experimental heterogeneity need to be considered. Empirical and hierarchical Bayes methods offer a flexible modeling framework for such analyses, relying on empirical or subjective sources to determine the degree of pooling. These richer methods can be particularly useful to meta-analysis of experiments in parapsychology conducted under potentially diverse conditions.

For the recent ganzfeld series, assuming them to be independent binomially distributed as discussed in Section 5, the data can be summed (pooled) across series to estimate a common hit rate. Honorton et al. (1990) assessed the homogeneity of effects across the 11 series using a chi-square test that compares individual effect sizes to the weighted mean effect. The chi-square statistic $\chi^2_{10} = 16.25$, not statistically significant ($p = 0.093$), largely reflects the contribution of the last "special" series (contributes 9.2 units to the χ^2_{10} value), and to a lesser extent the novice series with a negative effect (contributes 2.5 units). The outlier series can be dropped from the analysis to provide a more conservative estimate of the presence of psi effects for this data (this result is reported in Section 5). For the remaining 10 series, the chi-square value $\chi^2_9 = 7.01$ strongly favors homogeneity, although more than one-third of its value is due to the novice series (number 4 in Table 1). This pattern points to the potential usefulness of a richer model to accommodate series that may be distinct from the others. For the earlier ganzfeld data analyzed by Honorton (1985b), the appeal of a Bayes or other model that recognizes the heterogeneity across studies is clear cut: $\chi^2_{23} = 56.6$, $p = 0.0001$, where only those studies with common chance hit rate have been included (see Table 2).

Historic reliance on voting-count approaches to determine the presence of psi effects makes it natural to consider Bayes models that focus on the ensemble of experimental effects from parapsychological studies, rather than individual estimates. Recent work in parapsychology that compares effect sizes across studies, rather than estimating separate study effects, reinforces the need to examine this type of model. Louis (1984) develops Bayes and empirical Bayes methods for problems that consider the ensemble of parameter values to be the primary goal, for example, multiple comparisons. For the simple compound normal model, $Y_i \sim N(\theta_i, 1)$, $\theta_i \sim N(\mu, \tau^2)$, the standard Bayes estimates (posterior means)

$$\theta_i^* = \mu + D(Y_i - \mu) \quad \text{and} \quad D = \frac{\tau^2}{1 + \tau^2}$$

where the θ_i represent experimental effects of interest, are modified approximately to

$$\theta_i^l \approx \mu + \sqrt{D}\,(Y_i - \mu)$$

when an ensemble loss function is assumed. The new estimates adjust the shrinkage factor D so that their sample mean and variance match the posterior expectation and variance of the θ's. Similar results are obtained when the model is gener-

TABLE 1
Recent ganzfeld series

Series type	N Trials	Hit rate	Y_i	σ_i
Pilot	22	0.36	−0.58	0.44
Pilot	9	0.33	−0.71	0.71
Pilot	36	0.28	−0.94	0.37
Novice	50	0.24	−1.15	0.33
Novice	50	0.36	−0.58	0.30
Novice	50	0.30	−0.85	0.31
Novice	50	0.36	−0.58	0.30
Novice	6	0.67	0.71	0.87
Experienced	7	0.43	−0.28	0.76
Experienced	50	0.30	−0.85	0.31
Experienced	25	0.64	0.58	0.42
Overall	355	0.34		

TABLE 2
Earlier ganzfeld studies

N Trials	Hit rate	Y_i	σ_i
32	0.44	−0.24	0.36
7	0.86	1.82	1.09
30	0.43	−0.28	0.37
30	0.23	−1.21	0.43
20	0.10	−2.20	0.75
10	0.90	2.20	1.05
10	0.40	−0.41	0.65
28	0.29	−0.90	0.42
10	0.40	−0.41	0.65
20	0.35	−0.62	0.47
26	0.31	−0.80	0.42
20	0.45	−0.20	0.45
20	0.45	−0.20	0.45
30	0.53	0.12	0.37
36	0.33	−0.71	0.35
32	0.28	−0.94	0.39
40	0.28	−0.94	0.35
26	0.46	−0.16	0.39
20	0.60	0.41	0.46
100	0.41	−0.36	0.20
40	0.33	−0.71	0.34
27	0.41	−0.36	0.39
60	0.45	−0.20	0.26
48	0.21	−1.33	0.35
722	.38		

alized to the case of unequal variances, $Y_i \sim N(\theta_i, \sigma_i^2)$.

For the above model, the fraction of θ_i^l above (or below) a cut point C is a consistent estimate of the fraction of $\theta_i > C$ (or $\theta_i < C$). Thus, the use of ensemble, rather than component-wise, loss can help detect when individual effects are above a specified threshold by chance. For the meta-analysis of ganzfeld experiments, the observed binomial proportions transformed on the logit (or arcsin$\sqrt{}$) scale can be modeled in this framework. Letting d_i and m_i denote the number of direct hits and misses respectively for the ith experiment, and p_i as the corresponding population proportion of direct hits, the Y_i are the observed logits

$$Y_i = \log(d_i/m_i)$$

and σ_i^2, estimated by maximum likelihood as $1/d_i + 1/m_i$, is the variance of Y_i conditional on $\theta_i = \text{logit}(p_i)$. The threshold logit (0.25) ≈ 1.10 can be used to identify the number of experiments for which the proportion of direct hits exceeds that expected by chance.

Table 1 shows Y_i and σ_i for the 11 ganzfeld series. All but one of the series are well above the threshold; Y_4 marginally falls below −1.10. Any shrinkage toward a common hit rate will lead to an estimate, θ_4^* or θ_4^l, above the threshold. The use of ensemble loss (with its consistency property) provides more convincing support that all $\theta_i > -1.10$, although posterior estimates of uncertainty are needed to fully calibrate this. For the earlier ganzfeld data in Table 2, ensemble loss can similarly be used to determine the number of studies with $\theta_i < -1.10$ and specifically whether the negative effects of studies 4 and 24 ($Y_4 = -1.21$ and $Y_{24} = -1.33$) occurred as a result of chance fluctuation.

Features of the ganzfeld data in Section 5, such as the outlier series, suggest that further elaboration of the basic Bayesian set-up may be necessary for some meta-analyses in parapsychology. Hierarchical models provide a natural framework to specify these elaborations and explore how results change with the prior specification. This type of sensitivity analysis can expose whether conclusions are closely tied to prior beliefs, as observed by Jeffreys for RNG data (see Section 7). Quantifying the influence of model components deemed to be more subjective or less certain is important to broad acceptance of results as evidence of psi performance (or lack thereof).

Consider the initial model commonly used for Bayesian analysis of discrete data:

$$Y_i \mid p_i, n_i \sim B(p_i, n_i),$$
$$\theta_i \sim N(\mu, \tau^2), \quad \theta_i = \text{logit}(p_i),$$

with noninformative priors assumed for μ and τ^2 (e.g., log τ locally uniform). The distinctiveness of the last "special" series and, in general, the different types of series (pilot versus formal, novice versus experienced) raises the question of whether the experimental effects follow a normal distribution. Weighted normal plots (Ryan and Dempster, 1984) can be used to graphically diagnose the adequacy of second-stage normality (see Dempster, Selwyn and Weeks, 1983, for examples with binary response and normal superpopulation).

Alternatively, if nonnormality is suspected, the model can be revised to include some sort of heavy-tailed prior to accommodate possibly outlying series or studies. West (1985) incorporates additional scale parameters, one for each component of the model (experiment), that flexibly adapt to a typical θ_i and discount their influence on posterior estimates, thus avoiding under- or over-shrinkage due to such θ_i. For example, the second stage can specify the prior as a scale mixture of normals:

$$\theta_i \sim N(\mu, \tau^2\gamma_i^{-1}),$$
$$k\gamma_i \sim \chi_k^2,$$
$$v\tau^{-2} \sim \chi_v^2.$$

This approach for the prior is similar to others for

maximum likelihood estimation that modify the sampling error distribution to yield estimates that are "robust" against outlying observations.

Like its maximum likelihood counterparts, in addition to the robust effect estimates θ_i^*, the Bayes model provides (posterior) scale estimates γ_i^*. These can be interpreted as the weight given to the data for each θ_i in the analysis and are useful to diagnosing which model components (series or studies) are unusual and how they influence the shrinkage. When more complex groupings among the θ_i are suspected, for example, bimodal distribution of studies from different sites or experimenters, other mixture specifications can be used to further relax the shrinkage toward a common value.

For the 11 ganzfeld series, the last "outlier" series, quite distinct from the others (hit rate = 0.64), is moderately precise ($N = 25$). Omitting it from the analysis causes the overall hit rate to drop from 0.344 to 0.321. The scale mixture model is a compromise between these two values (on the logit scale), discounting the influence of series 11 on the estimated posterior common hit rate used for shrinkage. The scale factor γ_{11}^*, an indication of how separate θ_{11} is from the other parameters, also causes θ_{11}^* to be shrunk less toward the common hit rate than other, more homogeneous θ_i, giving more weight to individual information for that series (see West, 1985). The heterogeneity of the earlier ganzfeld data is more pronounced, and studies are taken from a variety of sources over time. For these data, the γ_i^* can be used to explore atypical studies (e.g., study 6, with hit rate = 0.90, contributes more than 25% to the χ_{23}^2 value for homogeneity) and groupings among effects, as well as protect the analysis from misspecification of second-stage normality.

Variation among ganzfeld series or studies and the degree to which pooling or shrinking is appropriate can be investigated further by considering a range of priors for τ^2. If the marginal likelihood of τ^2 dominates the prior specification, then results should not vary as the prior for τ^2 is varied. Otherwise, it is important to identify the degree to which subjective information about interexperimental variability influences the conclusions. This sensitivity analysis is a Bayesian enrichment of the simpler test of homogeneity directed toward determining whether or not complete pooling is appropriate.

To assess how well heterogeneity among historical control groups is determined by the data. Dempster, Selwyn and Weeks (1983) propose three priors for τ^2 in the logistic-normal model. The prior distributions range from strongly favoring individual estimates, $p(\tau^2)d\tau \propto \tau^{-1}$, to the uniform reference prior $p(\tau^2)d\tau \propto \tau^{-2}$, flat on the log τ scale, to strongly favoring complete pooling, $p(\tau^2)d\tau \propto \tau^{-3}$ (the latter forcing complete pooling for the compound normal model; see Morris, 1983). For their two examples, the results (estimates of linear treatment effects) are largely insensitive to variation in the prior distribution, but the number of studies in each example was large (70 and 19 studies available for pooling). For the 11 ganzfeld series, τ^2 may be less well determined by the data. The posterior estimate of τ^2 and its sensitivity to $p(\tau^2)d\tau$ will also depend on whether individual scale parameters are incorporated into the model. Discounting the influence of the last series will both shift the marginal likelihood toward smaller values of τ^2 and concentrate it more in that region.

The issue of objective assessment of experiment results is one that extends well beyond the field of parapsychology, and this paper provides insight into issues surrounding the analysis and interpretation of small effects from related studies. Bayes methods can contribute to such meta-analyses in two ways. They permit experimental and subjective evidence to be formally combined to determine the presence or absence of effects that are not clear cut or controversial (e.g., psi abilities). They can also help uncover sources and degree of uncertainty in the scientific conclusions.