Biometrika (1997), **84**, 2, *pp.* 419–427
Printed in Great Britain

A Bayesian perspective on the Bonferroni adjustment

BY PETER H. WESTFALL
Department of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, Texas 79409, U.S.A.
e-mail: westfall@ttu.edu

WESLEY O. JOHNSON AND JESSICA M. UTTS
Division of Statistics, University of California at Davis, Davis, California 95616, U.S.A.
e-mail: wojohnson@ucdavis.edu jmutts@ucdavis.edu

SUMMARY

Bayes/frequentist correspondences between the p-value and the posterior probability of the null hypothesis have been studied in univariate hypothesis testing situations. This paper extends these comparisons to multiple testing and in particular to the Bonferroni multiple testing method, in which p-values are adjusted by multiplying by k, the number of tests considered. In the Bayesian setting, prior assessments may need to be adjusted to account for multiple hypotheses, resulting in corresponding adjustments to the posterior probabilities. Conditions are given for which the adjusted posterior probabilities roughly correspond to Bonferroni adjusted p-values.

Some key words: Adjusted p-value; Bayes factor; Multiple testing; Multiplicity adjustment; Simultaneous inference.

1. INTRODUCTION

Multiple testing is difficult and controversial on either side of the Bayes/frequentist fence, with arguments over whether and how multiplicity adjustment should be performed. Nevertheless, the issue is important since practising statisticians routinely deal with large and complex datasets, with multiple hypotheses of interest.

We describe multiple testing situations for which the Bonferroni correction and Bayesian analysis roughly coincide. This happens when (i) the hypotheses tested are regarded independently, (ii) there is a concern that many or all of the hypotheses tested might be true, and (iii) there is interest in the individual hypotheses, rather than a single omnibus test. These are precisely the conditions where Bonferroni-style adjustments are typically applied in practice. As in the univariate case, the reconciliation of Bayesian probabilities and frequentist p-values is tenuous in the multiple testing case. Nevertheless, we identify situations where multiplication by the factor k, as is done with the Bonferroni adjustment, has Bayesian rationale.

In § 2 we review Bayesian measures of evidence and the p-value. In § 3, multiple testing concepts are reviewed from the frequentist perspective, and § 4 compares frequentist multiplicity adjustments and revised Bayesian posterior probabilities, with applications to independent tests, correlated tests and pairwise comparisons. An example is given in § 5 and concluding remarks are given in § 6. Details of our prior elicitation for the example are given in an Appendix.

2. UNIVARIATE TESTS

Many authors have compared frequentist p-values, p^F, and Bayesian measures of evidence, as measured by the Bayesian probability $p^B = \text{pr}(H_0|\text{the data})$, finding various degrees of reconciliation depending upon the type of test considered. Casella & Berger (1987) considered the one-sided testing problem, Berger & Sellke (1987) addressed the two-sided testing problem, and Moreno & Cano (1989) extended the results to multidimensional point null tests. In these papers, and in ours, the Bayes factor, b, is used. When the null is $H_0: \theta = \theta_0$,

$$b = \frac{f(z|\theta_0)}{\int f(z|\theta)g(\theta)\,d\theta},$$

where $f(z|\theta)$ is the probability density function of z and $g(\theta)$ is the prior density for θ. If $\pi_0 = \text{pr}(H_0)$ is specified, the posterior probability is

$$p^B = \left\{1 + \frac{(1-\pi_0)}{\pi_0}\frac{1}{b}\right\}^{-1}. \tag{1}$$

Kass & Raftery (1995) review issues surrounding the use and calculation of Bayes factors. We assume that all Bayes factors are known, and consider the effects of varying the assessment $\text{pr}(H_0)$.

3. FREQUENTIST MULTIPLE TESTING

Consider a collection of p-values $\{p_i^F; i = 1, \ldots, k\}$ corresponding to null hypotheses H_{0i} $(i = 1, \ldots, k)$. The usual rejection rule $(p_i^F \leqslant 0{\cdot}05)$ can be criticised on the grounds that if many or all of the null hypotheses tested are true, some of them will be incorrectly rejected and the familywise error rate can be much larger than 0·05. Multiple testing methods commonly aim to control this error rate at a pre-specified level; Hochberg & Tamhane (1987) provide a good general reference.

From the frequentist standpoint, a reasonable measure of evidence concerning a particular hypothesis is the 'adjusted p-value', $\tilde{p}_i^F$, which is the smallest familywise error rate for which H_{0i} would be rejected. The simple Bonferroni method leads to adjusted p-values $\tilde{p}_i^F = kp_i^F$, which are usually upper bounds on the required amount of adjustment in frequentist multiple testing. Holm (1979), Shaffer (1986), Hochberg (1988), Dunnett & Tamhane (1992) and Westfall & Young (1993) give improvements over the simple Bonferroni method. Nevertheless, the Bonferroni adjustment is a convenient generic reference point to facilitate our comparison with Bayesian methods.

4. WHEN IS BAYES LIKE BONFERRONI?

4·1. *Interval nulls and shrinkage priors*

From the Bayesian point of view, there is no need to adjust the posterior probability of the event $\{H_{0i}$ is true$\}$ provided one's prior is well calibrated. Recognising that extreme statistics can and do occur in multiple testing applications, Bayesians commonly consider priors that 'shrink' the observed effects toward some common mean (Meng & Dempster, 1987; Berry, 1988; Berger & Deely, 1988; Lindley, 1990).

A typical 'shrinkage' prior specifies normal and exchangeable θ_i's with mean 0, variance σ^2 and correlation ρ. Suppose the observable data vector is $Z = (Z_1, \ldots, Z_k)$, the parameter

vector is $\Theta = (\theta_1, \ldots, \theta_k)$, and Z given Θ is distributed as $N_k(\Theta, I)$. The assumption of conditional independence is valid in many multiple testing applications (Proschan & Follmann, 1995); a violation occurs when the θ_i represent pairwise contrasts among the means of treated groups.

The posterior distribution of Θ is multivariate normal with marginals satisfying

$$E(\theta_i|z) = z_i \frac{\gamma}{1+\gamma} + \frac{\bar{z}}{(1+\gamma)^2/(k\rho\sigma^2) + 1 + \gamma}, \tag{2}$$

$$\text{var}(\theta_i|z) = \frac{\gamma}{1+\gamma} + \frac{1}{(1+\gamma)^2/(\rho\sigma^2) + k + k\gamma}, \tag{3}$$

$$\text{corr}(\theta_i, \theta_j|z) = \frac{\rho}{1 + \gamma^2/\sigma^2 + k\rho\gamma}, \tag{4}$$

where $\gamma = (1-\rho)\sigma^2$. With large k, (2) and (3) reduce to $\{\gamma z_i + \bar{z}\}/\{1+\gamma\} + o(1)$ and $\gamma/(1+\gamma) + o(1)$, respectively.

Consider testing one-sided hypotheses $H_{0i}: \theta_i \leqslant 0$ versus $H_{1i}: \theta_i > 0$. In this case the Bayesian probability $\text{pr}(H_{0i}|z) = \text{pr}(\theta_i \leqslant 0|z)$ and the Bonferroni p-value $\tilde{p}_i^F = kp_i^F$ cannot be reconciled using the shrinkage prior; we have $k^{-1}\tilde{p}_i^F = p_i^F$, which is constant in k for every $z \in \mathscr{R}^k$, while $k^{-1}\,\text{pr}(H_{0i}|z)$ tends to zero for every $z \in \mathscr{R}^k$ as k tends to infinity.

This nonreconciliation is not surprising, since the given prior implies

$$(0{\cdot}5)^k \leqslant \text{pr}(H_{0i} \text{ is true, all } i) \leqslant 0{\cdot}5,$$

with extremes occurring in the cases (a) complete independence of the θ_i ($\rho = 0$), or (b) perfect correlation among the θ_i ($\rho = 1$). In the case of near or complete independence, the event $\{H_{0i}$ is true, all $i\}$ is not considered likely. The Bonferroni method is based upon the implicit presumption of a moderate degree of belief in the event $\{H_{0i}$ is true, all $i\}$; therefore, it is not surprising that the frequentist and Bayesian methods differ when this event is explicitly considered to be a priori implausible.

4·2. *Point nulls and Bayes factors*

The implicit motivation for the Bonferroni correction is a concern that the event $\{H_{0i}$ is true, all $i\}$ is plausible, with probability perhaps in a neighbourhood of 0·5, and certainly not arbitrarily close to zero. In the Bayesian setting, suppose $\text{pr}(H_{0i}$ is true) is initially set at 0·5 for each i because the assessor is truly objective about the truth or falsity of the hypotheses. If the hypotheses are considered to be independent, then the prior probability of the event that all hypotheses are true is $(0{\cdot}5)^k$, which may actually be much smaller than the assessor intended for the joint probability. In this case the prior is not well calibrated and an adjustment is necessary. If, for example, $\text{pr}(H_{0i}$ is true, all $i) =$ 0·5 seems more reasonable, and if there is no preference for particular H_{0i}, then the marginal probabilities $\text{pr}(H_{0i}$ is true) should be revised to $(0{\cdot}5)^{1/k}$. If the revised prior is used, the posterior probabilities $\text{pr}(H_{0i}|z)$ are also revised. In this setting, when k is large and b is small we obtain the following result; like the Bonferroni corrections, the Bayesian revised posterior probabilities are approximately proportional to $k \times$ (the original posterior probabilities). Notice that this is precisely the situation of concern in multiple testing; large k and small b imply that there are a large number of tests conducted and the data tend to support some alternative hypotheses.

Let the Bayes factor for H_{0i} be b_i. If $\text{pr}(H_{0i}$ is true) $= \pi_{0i}$ then p_i^B, the posterior probability

of H_{0i}, is given by (1). Suppose however that the product of the π_{0i}, pr(H_{0i} is true, all i), is considered too small. Then a revision of the prior is necessary. Suppose the analyst believes that pr(H_{0i} is true, all i) $= \Pi_0$. Let $\tilde{\pi}_{0i}$ be the revised prior probability of H_{0i}, chosen so that the product of the $\tilde{\pi}_{0i}$ is Π_0. One possibility, if the H_{0i} are exchangeable, is to set $\tilde{\pi}_{0i} \equiv \tilde{\pi}_0 = \Pi_0^{1/k}$, as suggested by Dawid (1987), who examined multiple hypotheses in court cases.

Let $\tilde{p}_i^B$ be the revised posterior probability for H_{0i}. The following proposition gives conditions under which $\tilde{p}_i^B \propto kp_i^B$, similar to the Bonferroni correction. We assume for simplicity that the marginal probabilities are identical, and therefore drop the subscript i. We also assume that the $kb_i \to d \geqslant 0$, which implies that the Bayes factor has order k^{-1} or smaller.

PROPOSITION 1. *Assume that the H_{0i} are exchangeable and $\tilde{\pi}_{0i} \equiv \tilde{\pi}_0 = \Pi_0^{1/k}$. Let $k \to \infty$ and $b_i \to 0$ in such a way that $kb_i \to d \geqslant 0$. Then*

$$\lim_{k\to\infty} \frac{kp_i^B}{\tilde{p}_i^B} = \{d - \ln(\Pi_0)\} \frac{\pi_0}{1-\pi_0}.$$

Proof. Note that

$$\frac{kp_i^B}{\tilde{p}_i^B} = \frac{\{\Pi_0^{1/k} kb_i + k(1-\Pi_0^{1/k})\}\pi_0}{(\pi_0 b_i + 1 - \pi_0)\Pi_0^{1/k}}.$$

The result follows since $b_i \to 0$, $kb_i \to d$, $\Pi_0^{1/k} \to 1$ and $k(1-\Pi_0^{1/k}) \to -\ln(\Pi_0)$. □

Thus the Bayesian revised probability $\tilde{p}_i^B$ is approximately equal to the original Bayes probability p_i^B multiplied by ck, where $c^{-1} = \{d - \ln(\Pi_0)\}\pi_0(1-\pi_0)^{-1}$. This approximation is best when b_i is small, the case of most common concern. For example, if the marginal prior probabilities $\pi_0 = 0{\cdot}5$ are revised to $\tilde{\pi}_0 = (0{\cdot}5)^{1/k}$ so that $\Pi_0 = 0{\cdot}5$, and if $kb_i \doteq 0$, then the multiplier is $k/\ln(2) \doteq 1{\cdot}4k$. The frequentist adjustment is similar, since the adjusted p-value is the original p-value multiplied by k.

The actual Bonferroni multiplier k is obtained in a variety of ways. One way is shown in the following result.

COROLLARY. *Let $\tilde{\pi}_0 = 1 - 1/k$, $k \to \infty$ and $b_i \to 0$ in such a way that $kb_i \to 0$, and let $\pi_0 = 0{\cdot}5$. Then $kp_i^B/\tilde{p}_i^B \to 1$.*

Proof. The proof follows by substituting $1 - 1/k$ for $\Pi_0^{1/k}$ in the proof of Proposition 1. □

This condition on $\tilde{\pi}_0$ is essentially equivalent to taking $\Pi_0 = e^{-1}$.

4·3. *Dependent hypotheses*

The Bayesian adjustment shown in Proposition 1 is less extreme with dependent hypotheses, as with frequentist multiplicity adjustments. When the H_{0i} are dependent, the premise $\tilde{\pi}_0 = \Pi_0^{1/k}$ is unrealistic. An example is the case of pairwise comparison of t means $\mu_1, \ldots, \mu_t$. Here, the θ_i are more properly doubly-indexed,

$$\theta_{ij} = \mu_i - \mu_j \quad (i<j), \quad k = \binom{t}{2}.$$

Clearly, the k hypotheses $H_{0ij}: \theta_{ij} = 0$ cannot be considered to be independent events. To

model the dependence structure, consider the following hierarchical prior for the μ_i: given a hyperparameter μ with continuous prior F, suppose that the μ_i are conditionally independent with

$$\mu_i|\mu \begin{cases} \equiv \mu & \text{with probability } \lambda, \\ \sim G_i & \text{with probability } 1-\lambda, \end{cases}$$

for continuous priors G_i. Then $\text{pr}(\theta_{ij}=0)=\lambda^2$ and $\text{pr}(\text{all } \theta_{ij}=0)=\lambda^t$.

Proposition 1 is then applicable as follows. Suppose the initial specification of the marginal priors has $\pi_0=\lambda^2$. If the joint prior is thought to be $\text{pr}(\text{all } \theta_{ij}=0)=\Pi_0$, then the revised value for λ is $\tilde{\lambda}=\Pi_0^{1/t}$ or equivalently $\tilde{\pi}_0=(\Pi_0^2)^{1/t}$, and the result of Proposition 1 follows by substituting t for k and Π_0^2 for Π_0. Since t is roughly proportional to $k^{\frac{1}{2}}$, we see a less severe adjustment in this application, due to the dependence structure among tests. The Bonferroni multiplier $k^{\frac{1}{2}}$ also has been suggested by John Tukey, as reported in Mantel (1980), to correct for multiplicity with dependent tests.

As a second example, consider the exchangeable hypotheses of § 4·1, and consider calibrating the prior so that $\text{pr}(\theta_i \leqslant 0, \text{ all } i)=\Pi_0$. For $\rho=0$, we find $\Pi_0=(0{\cdot}5)^k$ and, for $\rho=1$, we find $\Pi_0=0{\cdot}5$. We now show that, in general, $\text{pr}(\theta_i \leqslant 0, \text{ all } i)$ is completely determined by the choice of ρ, in that one or the other of these may be specified, but the other is then determined. Further, if $\rho \geqslant 0$, the shrinkage prior requires that $(0{\cdot}5)^k \leqslant \Pi_0 \leqslant 0{\cdot}5$, so that, if one's beliefs dictate that $\Pi_0 > 0{\cdot}5$, a different prior must be selected.

Let $\theta_1, \ldots, \theta_k$ given μ, be a random sample from $N(\mu, \sigma_e^2)$ and assume $\mu \sim N(0, \sigma_m^2)$. Then $\theta_1, \ldots, \theta_k$ are marginally exchangeable and normal with mean zero, variance $\sigma_m^2+\sigma_e^2=\sigma^2$ and with correlation $\rho=\sigma_m^2/\sigma^2$. Then

$$\text{pr}(\theta_i \leqslant 0, \text{ all } i) = E\{\text{pr}(\theta_i \leqslant 0, \text{ all } i|\mu)\} = \int \Phi^k \left\{ -z\left(\frac{\rho}{1-\rho}\right)^{\frac{1}{2}} \right\} \frac{e^{-z^2/2}}{(2\pi)^{\frac{1}{2}}}\, dz, \qquad (5)$$

where $\Phi(.)$ is the cumulative standard normal distribution function. Since (5) is monotone in ρ, it is a simple matter to find the value of ρ that results in one's calibrated choice for Π_0. For instance, using numerical integration, we found that, for $k=5$, 10 and 25, when $\Pi_0=0{\cdot}4$, ρ was 0·954, 0·973 and 0·984, respectively, which is not surprising since, when $\Pi_0=0{\cdot}5$, $\rho=1$ for all k. For $k=5$, 10 and 25, when $\Pi_0=0{\cdot}1$, ρ is 0·284, 0·526 and 0·678.

The posterior for $\theta_1, \ldots, \theta_k$ is also normal, with the same correlation structure as the prior, see (3) and (4), but with nonzero means. Integration of the resulting multivariate normal is straightforward for determining one's posterior belief, $\text{pr}(\theta_i \leqslant 0, \text{ all } i|z)$.

The two special cases just considered make it clear that Bayesian multiplicity adjustments for dependent hypotheses are determined by the nature of the dependence. There is no simple adjustment that will work for all situations as Bonferroni does in the frequentist realm. Nonetheless, it is important to try to calibrate the individual prior assessments with the overall assessment, as we have done for these special cases.

5. Example

In the Summer of 1995, at the request of the U.S. Congress, one of us (Utts, 1995) was asked to evaluate a recently declassified government-sponsored research program into extrasensory perception (ESP). The majority of the experiments used 'remote viewing', in which a photograph or short video segment, denoted as the 'target', was randomly selected

from a larger set and displayed in one location, and a 'remote viewer' at a distant location was asked to provide a description. A question of interest was whether or not each of a specially selected group of remote viewers showed ESP ability. The government was more interested in individual abilities for 'intelligence gathering' than in proving the general existence of ESP.

Here, we use the accumulated data from two experiments, involving five remote viewers contributing between 40 and 70 trials each. Details of the experiments were published by May, Spottiswoode & James (1994) and by Lantz, Luke & May (1994). For each viewer, the null hypothesis is that he does not exhibit ESP. There are $k = 5$ hypotheses.

For each viewing, a 'blind judge' compared the viewer's written material with five potential targets, each of which could originally have been selected as the actual target with equal probability. The judge assigned a rank to each of the five choices, from 1 = best match to 5 = worst, and the rank assigned to the real target was the 'score' for that guess.

Let R_{ij} be the rank assigned to the actual target for viewer i and trial j. Under the null hypotheses $E(R_{ij}) = 3$ and $\text{var}(R_{ij}) = 2$. When n_i is large, a one-sided frequentist test can be conducted by constructing z-scores for each viewer; $z_i = (n_i/2)^{\frac{1}{2}}(\bar{R}_i - 3)$. The results are in Table 1.

With the standard rejection criterion at the 5% level, the null hypothesis is rejected for viewers 1, 2 and 5. A Bonferroni correction requires $p \leqslant 0{\cdot}01$, but two viewers still provide evidence of ESP under this criterion.

For the Bayesian tests, Dr Edwin May provided information to construct priors for the alternative hypotheses. He termed viewers 1 and 2 highly skilled and the other three skilled, and we constructed separate priors for the two groups. Unavoidably, Dr May had seen the data before we questioned him, but he assured us that after many years of working with these viewers his assessment would have led to the same prior. As described briefly in the Appendix, we used his prior assessments regarding $(p_1, \ldots, p_5)$, where

$$p_i = \text{pr}(\text{actual target is given rank } i),$$

to induce a prior distribution on the effect sizes expected for each group; see Fig. 1. In this context an effect size is given by $\theta_i \equiv (\delta_i - 3)/2^{\frac{1}{2}}$, where $\delta_i = E(R_{ij}) = \sum_{i=1}^{5} ip_i$. Notice that θ_i is estimated by $z_i/n_i^{1/2}$.

We wished to test the hypotheses $H_{0i}: \theta_i = 0$ versus $H_{1i}: \theta_i < 0$. The Bayes factors in Table 1 were calculated using simple Monte Carlo integration under the assumption that $z_i/n_i^{1/2}$ given θ_i is approximately normal, and that θ_i has the approximate prior distribution given in Fig. 1.

We computed posterior probabilities for the alternative hypotheses for low, medium and high prior expectations for individual viewers, respectively $\text{pr}(H_{0i}) = 0{\cdot}9$, $\text{pr}(H_{0i}) = 0{\cdot}5$

Table 1. *Extrasensory perception example: statistics for viewers*

	Viewer				
	1	2	3	4	5
n_i	60	70	40	60	60
$\bar{R}_i$	2·45	2·47	2·80	2·93	2·62
z_i	−3·01	−3·14	−0·89	−0·38	−2·08
p-value	0·0013	0·0008	0·1867	0·3520	0·0188
Bayes factor	0·0177	0·0139	1·1837	4·4978	0·2242

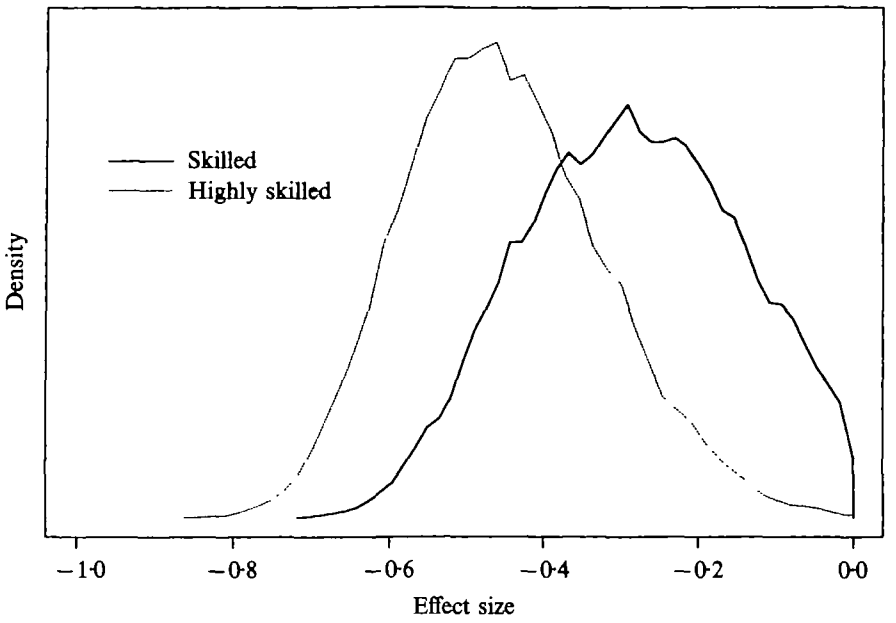

Fig. 1. Extrasensory perception example: induced priors for effect size from data in Table 1. Viewers 1 and 2, highly skilled; viewers 3–5, skilled.

and $\text{pr}(H_{0i}) = 0{\cdot}1$. If all five null hypotheses are regarded independently, the resulting low, medium and high overall probabilities are 0·59, 0·03 and 0·00001, respectively, which may be smaller than desired for each type of belief, thus requiring an adjustment.

Suppose in each case we adjust π_0 to be $\tilde{\pi}_0 = \pi_0^{1/5}$ as if the initial degree of belief in $\text{pr}(H_{0i}$ is true) is meant to apply to $\text{pr}(H_{0i}$ is true, all i). This results in revised probabilities $\tilde{\pi}_0$ of 0·979, 0·87 and 0·63 for low, medium and high prior expectations, respectively. Table 2 shows posterior probabilities for each type of prior before and after adjusting them for multiple testing, as well as the ratio of these probabilities. Notice that the adjustment is most relevant when the Bayes factor b_i is small. This occurs for viewers 1 and 2, and to a lesser extent for viewer 5, and is similar to a Bonferroni adjustment in this case for the medium prior expectation. Note also that multiplicative adjustments can get quite large with small b_i, as suggested by Proposition 1. The approximate number given by Proposition 1 for small b_i, $k\{-\ln(\Pi_0)\}\{\pi_0/(1-\pi_0)\}^{-1}$, results in the values 5·27, 7·21 and 19·54 for low, medium and high prior expectations. These values most closely reflect the pattern of the observed multipliers in Table 2 for viewer 2, who had the smallest Bayes factor. According to Proposition 1, the approximations will become closer for smaller b_i and larger k.

Table 2. *Bayes factors and posterior probabilities*

Viewer	Bayes factor	Prior $\pi_0 = 0{\cdot}9$, $\tilde{\pi}_0 = 0{\cdot}979$			Prior $\pi_0 = 0{\cdot}5$, $\tilde{\pi}_0 = 0{\cdot}87$			Prior $\pi_0 = 0{\cdot}1$, $\tilde{\pi}_0 = 0{\cdot}63$		
		p_i^B	$\tilde{p}_i^B$	$\tilde{p}_i^B/p_i^B$	p_i^B	$\tilde{p}_i^B$	$\tilde{p}_i^B/p_i^B$	p_i^B	$\tilde{p}_i^B$	$\tilde{p}_i^B/p_i^B$
1	0·01774	0·1377	0·4545	3·30	0·0174	0·1066	6·11	0·0020	0·0294	14·96
2	0·01389	0·1111	0·3948	3·55	0·0137	0·0854	6·24	0·0015	0·0232	15·05
3	1·18374	0·9142	0·9823	1·07	0·5421	0·8884	1·64	0·1162	0·6693	5·76
4	4·49783	0·9759	0·9953	1·02	0·8181	0·9680	1·18	0·3332	0·8849	2·66
5	0·22417	0·6686	0·9132	1·37	0·1831	0·6012	3·28	0·0243	0·2771	11·40

6. DISCUSSION

The revision $\tilde{\pi}_0 = \Pi_0^{1/k}$ seems to suggest that pr(H_{0i} is true) should be revised as more tests are considered in the context of the same experiment. This is not the case. Marginal priors for individual hypotheses should not depend upon how many tests are examined; rather, prior assessments should be made on a study-by-study basis. The careful assessment of the joint probability on all null hypotheses can only improve the calibration of the joint prior. If the joint prior probability on all null hypotheses is assessed to be moderate, and if the prior correlation between the parameters is small, then we are forced to conclude that the individual prior probability on each null hypothesis is large, perhaps much larger than 0·5.

One also might think that $\tilde{\pi}_0 = \Pi_0^{1/k}$ is generally 'too large'. However, Box & Meyer (1986) use pr(H_{0i} is true) $\equiv 0{\cdot}8$ when selecting active experimental effects; Garthwaite & Dickey (1992) elicit priors from an expert as high as pr(H_{0i} is true) $= 0{\cdot}9$; and Dawid (1987) gives a similar construction, $\pi_0 = \Pi_0^{1/k}$, where the individual hypotheses refer to 'component issues' in a litigation, and where the collection of null hypotheses refers to the 'conjunction' of all such component issues.

For those who desire a Bayesian rationale, our results suggest that Bonferroni-style multiplicity adjustment may be appropriate, provided (a) it is thought that, simultaneously, all null hypotheses are moderately probable, and (b) the prior dependence between null hypotheses is small. Further, the Bayesian correspondences suggest that frequentists could account for a priori knowledge in an informal fashion as follows. First, do not multiplicity-adjust a test when the null hypothesis is suspected to be false a priori. For such a test, the prior $\pi_0 = \Pi_0^{1/k}$ will be too large. Secondly, after removing such suspected hypotheses, perform multiplicity adjustment on the remaining tests, provided that the truth of all these hypotheses is believed moderately probable.

This protocol provides an answer to the question 'why not perform multiplicity adjustments for all tests considered in the statistician's lifetime?' The answer is because it is quite unlikely that all null hypotheses considered in a lifetime will actually be true. By analogy, Kadane (1987) has claimed that it seems silly to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ if it is not thought remotely possbile that $\theta = 0$.

The primary message of this paper for Bayesians is that they should calibrate their priors according to their beliefs about pr(H_{0i} is true, all i): their marginal probabilities pr(H_{0i} is true) might be much larger than they had thought.

We suggest the following Bayesian protocol.

(i) Assign larger prior probabilities to null hypotheses that are not suspected to be false, a priori, than to those that are suspected to be false.

(ii) Explicitly model prior dependencies among hypotheses that are related, as shown, for example, in § 4·3.

(iii) Assign a value to pr(H_{0i} is true, all i). Use this probability, as well as correlational information, to assign priors on individual null hypotheses.

Frequentist multiplicity adjustments have been criticised by Bayesians (Berry, 1988; Lindley, 1990). The analyses of this paper suggest that frequentist and Bayesian multiple testing analyses need not be grossly disparate.

APPENDIX

Prior elicitation for the extrasensory perception example

To determine a prior on the effect sizes for the two types of viewer, we first elicited a prior distribution on the vector of probabilities $(p_1, p_2, \ldots, p_5)$. We did not simply take a Dirichlet prior

on this vector because we wanted to give the most weight to the information for direct hits and so on, with the least weight for the information about a rank of 5.

For each i, we asked Dr May to provide a value for the 50th and 99th percentiles for his assessment of the probability with which each viewing would get a rank of i, given that it did not receive a rank less than i. We then used the beta density with matching 50th and 99th percentiles as the prior for the conditional density of p_i, given $p_1, \ldots, p_{i-1}$. For the highly skilled viewers, he was 99% sure that the probability of a direct hit, p_1, would be 0·4 or less, and 50% sure it would be 0·3 or less, resulting in a beta (52, 78) prior for p_1. Similarly, his assessment of the probability of a second place match, given that a direct hit was not obtained, resulted in a beta (24, 28) conditional density for p_2.

We repeated this for all i and for both types of viewer, inducing joint prior probability distributions on $(p_1, p_2, \ldots, p_5)$. We then used the prior to generate the density estimates for the effect size, shown in Fig. 1, using 10 000 Monte Carlo samples for each type of viewer. Since Dr May would truncate the possibilities at effect sizes of zero, we discarded any sample for which the effect size was greater than 0, and renormalised.

References

Berger, J. O. & Deely, J. (1988). A Bayesian approach to ranking and selection of related means with alternatives to analysis-of-variance methodology. *J. Am. Statist. Assoc.* **83**, 364–73.

Berger, J. O. & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence (with comments). *J. Am. Statist. Assoc.* **82**, 112–22.

Berry, D. A. (1988). Multiple comparisons, multiple tests, and data dredging: A Bayesian perspective. In *Bayesian Statistics 3*, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 79–94. Oxford University Press.

Box, G. E. P. & Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics* **28**, 11–8.

Casella, G. & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with comments). *J. Am. Statist. Assoc.* **82**, 106–11.

Dawid, A. P. (1987). The difficulty about conjunction. *Statistician* **36**, 91–7.

Dunnett, C. W. & Tamhane, A. C. (1992). A step-up multiple test procedure. *J. Am. Statist. Assoc.* **87**, 162–70.

Garthwaite, P. H. & Dickey, J. M. (1992). Elicitation of prior distributions for variable-selection problems in regression. *Ann. Statist.* **20**, 1697–719.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–2.

Hochberg, Y. & Tamhane, A. C. (1987). *Multiple Comparison Procedures.* New York: John Wiley.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

Kadane, J. B. (1987). Comment on 'Testing precise hypotheses'. *Statist. Sci.* **2**, 347–8.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–95.

Lantz, N. D., Luke, W. L. W. & May, E. C. (1994). Target and sender dependencies in anomalous cognition experiments. *J. Parapsychol.* **58**, 285–302.

Lindley, D. V. (1990). The 1988 Wald Memorial Lectures: The present position of Bayesian statistics. *Statist. Sci.* **5**, 44–89.

Mantel, N. (1980). Assessing laboratory evidence for neoplastic activity. *Biometrics* **36**, 381–99.

May, E. C., Spottiswoode, S. J. P. & James, C. L. (1994). Managing the target-pool bandwidth—possible noise reduction for anomalous cognition experiments. *J. Parapsychol.* **58**, 303–13.

Meng, C. Y. K. & Dempster, A. P. (1987). A Bayesian approach to the multiplicity problem for significance testing with binomial data. *Biometrics* **43**, 301–11.

Moreno, E. & Cano, J. A. (1989). Testing a point null hypothesis: Asymptotic robust Bayesian analysis with respect to the priors given on a subsigma field. *Int. Statist. Rev.* **57**, 221–32.

Proschan, M. A. & Follmann, D. A. (1995). Multiple comparisons with control in a single experiment versus separate experiments: Why do we feel differently? *Am. Statistician* **49**, 144–9.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Am. Statist. Assoc.* **81**, 826–31.

Utts, J. (1995). An assessment of the evidence for psychic functioning. In *An Evolution of Remote Viewing: Research and Applications*, Ed. M. D. Mumford, A. M. Rose and D. A. Goslin, pp. 3-2–3-42. Washington, DC: American Institutes for Research. Reprinted (1996) *J. Sci. Explor.* **10**(1), 3–30.

Westfall, P. H. & Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment.* New York: John Wiley.

[*Received August* 1995. *Revised September* 1996]