IMS Lecture Notes–Monograph Series
Optimality: The Third Erich L. Lehmann Symposium
Vol. 57 (2009) 326–332

DOI: 10.1214/09-LNMS5719

Bayesian Decision Theory for Multiple Comparisons

Charles Lewis[1] **and Dorothy T. Thayer**[2]

Fordham University and Educational Testing Service

Abstract: Applying a decision theoretic approach to multiple comparisons very similar to that described by Lehmann [*Ann. Math. Statist.* **21** (1950) 1–26; *Ann. Math. Statist.* **28** (1975a) 1–25; *Ann. Math. Statist.* **28** (1975b) 547–572], we introduce a loss function based on the concept of the false discovery rate (*FDR*). We derive a Bayes rule for this loss function and show that it is very closely related to a Bayesian version of the original multiple comparisons procedure proposed by Benjamini and Hochberg [*J. Roy. Statist. Soc. Ser. B* **57** (1995) 289–300] to control the sampling theory *FDR*. We provide the results of a Monte Carlo simulation that illustrates the very similar sampling behavior of our Bayes rule and Benjamini and Hochberg's procedure when applied to making all pair-wise comparisons in a one-way fixed effects analysis of variance setup with 10 and with 20 means.

Contents

1. Introduction

A previous paper by the authors [8] considered the application of Bayesian decision theory to the multiple comparisons problem for random effects designs, following the earlier work of Shaffer [10], Duncan [3], and Waller and Duncan [12]. In our paper, we demonstrated that the Bayes rule for a per-comparison "0-1" loss function controls a random effects version of the false discovery rate (*FDR*), thus supporting and extending Shaffer's [10] results.

A recent paper by Sarkar and Zhou [9] adopts a random effects setup very similar to that of our earlier paper. Rather than considering Bayes rules, they introduce a procedure that controls the random effects *FDR* discussed by us while maximizing the random effects per-comparison power rate that we had considered. This approach produces substantial power gains over other procedures (including ours), but it "declares even small differences significant when τ [the between-groups standard deviation] is large, thereby achieving [even] greater power than the unadjusted

[1]Fordham University and Educational Testing Service
[2]Educational Testing Service
AMS 2000 subject classifications: 62J15, 62C10, 62F15.
Keywords and phrases: Bayesian, decision theory, loss function, multiple comparisons, false discovery rate.

(per-comparison) procedure for large values of τ" (Sarkar & Zhou [9], p. 692). We view this as a weakness, rather than a strength, of their method, as it seems to ignore the basic principle behind multiple comparisons procedures, namely that making multiple inferences calls for increased conservatism relative to making a single inference.

The present study considers a more general setting for making multiple comparisons and introduces a new loss function that is more directly tied to the *FDR*. We derive a Bayes rule for this loss function and show that it is very closely related to a Bayesian version of the original multiple comparisons procedure proposed by Benjamini and Hochberg [1] to control the sampling theory *FDR*. We provide the results of a Monte Carlo simulation that illustrates the very similar sampling behavior of our Bayes rule and Benjamini and Hochberg's procedure when applied to testing all pairwise comparisons for a one-way fixed effects analysis of variance setup with 10 and with 20 means.

2. Setup

We start with a general likelihood $p(\mathbf{y}|\theta)$, prior $p(\theta)$, and resulting posterior $p(\theta|\mathbf{y})$. Let $\psi = \mathbf{f}(\theta)$ be a vector of m "contrasts" among the elements of θ. Suppose our goal is to identify the sign of each of the elements of ψ, given $\mathbf{y}$. In the language of decision theory, for each ψ_i, $i = 1, \ldots, m$, we will take action a_i, with $a_i = +1$ used to indicate that we declare ψ_i to be positive, $a_i = -1$ indicating that we declare ψ_i to be negative, and $a_i = 0$ used to indicate that we are unable to determine sign of ψ_i. Although directly inspired by Williams, Jones and Tukey [13], and Jones and Tukey [4], this approach to (multiple) hypothesis testing has its origins in the much earlier work of Lehmann [5–7].

To continue, we introduce two component loss functions: $L_1(\psi_i, a_i) = 1$ if the signs of ψ_i and a_i disagree, and $L_1 = 0$ otherwise (used to count the number of incorrect sign declarations); $L_2(\psi_i, a_i) = 1$ if $a_i = 0$, and $L_2 = 0$ otherwise (used to count the signs not declared). These actions and losses are very similar to those given by Lehmann ([7], p. 549). They differ from conventional treatments of hypothesis testing in the sense that they focus on identifying the sign of each contrast and do not formally consider the possibility that the value of the contrast could be (exactly) 0. The reasonableness of this approach, compared with conventional point hypothesis testing is emphasized by Jones and Tukey [4], among others.

We now propose a loss function that combines L_1 and L_2 as follows:

$$L_{DFDR}(\psi, \mathbf{a}) = \frac{\sum_{i=1}^{m} L_1(\psi_i, a_i)}{\max\{1, m - \sum_{i=1}^{m} L_2(\psi_i, a_i)\}} + \left(\frac{\alpha}{2}\right) \frac{\sum_{i=1}^{m} L_2(\psi_i, a_i)}{m}, \tag{1}$$

for a fixed choice of $0 < \alpha < 1$ (such as $\alpha = 0.05$). Here, *DFDR* (in notation introduced by Shaffer [11]) stands for Directional False Discovery Rate. The first term in equation (1) is the sample value of the *DFDR* for a given ψ and a vector of actions $\mathbf{a}$, namely the number of incorrect sign declarations, divided by the total number of signs declared (or divided by 1 if no signs are declared by $\mathbf{a}$).

The second term in equation (1) is $\alpha/2$ times the sample proportion of signs not declared. This term may be interpreted as a sample per-comparison Type II error rate, weighted by a relative importance factor of $\alpha/2$. Using a per-comparison formulation, as well as assigning this loss component a small weight, serves to emphasize that failure to declare a sign is considered to be much less serious than declaring that sign incorrectly. This emphasis is in keeping with the concern about controlling Type I errors (at the expense of making Type II errors) in traditional

treatments of the multiple comparisons problem. The Bayes decision rule for the complete loss function in equation (1) minimizes its posterior expected value, in this sense balancing the two types of losses against each other, with the major focus being on reducing the *DFDR*.

3. Bayes Decision Rule

To identify the actions that minimize the posterior expected value of the loss function given in equation (1), we begin by introducing some notation.

If $\Pr(\psi_i > 0|\mathbf{y}) > 0.5$, define $a_i^* = +1$ and $p_i = \Pr(\psi_i < 0|\mathbf{y})$; if $\Pr(\psi_i > 0|\mathbf{y}) \leq 0.5$, define $a_i^* = -1$ and $p_i = \Pr(\psi_i > 0|\mathbf{y})$. Note that a_i^* and p_i are related by the following result: $E_{\theta|\mathbf{y}}[L_1(\psi_i, a_i^*)|\mathbf{y}] = p_i$. Now order the p_i so that $p_{(1)} \leq \cdots \leq p_{(m)}$. Define $\mathbf{a}^{(k)}$ for $k = 1, \ldots, m$ as $a_{(i)}^{(k)} = a_{(i)}^*$, for $i = 1, \ldots, k$, and $a_{(i)}^{(k)} = 0$, for $i = k+1, \ldots, m$. For $k = 0$, take $a_{(i)}^{(0)} = 0$, for $i = 1, \ldots, m$.

The posterior expected loss for $\mathbf{a}^{(k)}$ is given by

$$E_{\theta|\mathbf{y}}\left[L_{DFDR}\left(\psi, \mathbf{a}^{(k)}\right)|\mathbf{y}\right] = \frac{\sum_{i=1}^{k} p_{(i)}}{\max\{1,k\}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k}{m}\right). \tag{2}$$

Clearly, $\mathbf{a}^{(k)}$ minimizes the posterior expected loss among all action vectors $\mathbf{a}$ that declare exactly k signs. Let k_{DFDR} be the value of k for which the posterior expected loss given in equation (2) is minimized. (The value of k_{DFDR} in a given setting would normally be determined by an exhaustive search over all values of $k = 0, \ldots, m$.) The Bayes decision rule for this problem is given by $\delta_{DFDR}(\mathbf{y}) = \mathbf{a}^{(k_{DFDR})}$, and the corresponding Bayes risk is

$$\begin{aligned} r\left(\delta_{DFDR}\right) &= E_{\theta,\mathbf{y}}\left[L_{DFDR}\left(\psi, \delta_{DFDR}\left(\mathbf{y}\right)\right)\right] \\ &= E_{\mathbf{y}}\left[\frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{\max\{1,k_{DFDR}\}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DFDR}}{m}\right)\right]. \end{aligned} \tag{3}$$

The latter expectation in equation (3) is taken with respect to the predictive distribution of $\mathbf{y}$, and it should be noted that the $p_{(i)}$ and, consequently, k_{DFDR} all depend on $\mathbf{y}$.

Since $E_{\theta|\mathbf{y}}[L_{DFDR}(\psi, \mathbf{a}^{(0)})|\mathbf{y}] = \alpha/2$ for all $\mathbf{y}$, it follows that $r(\delta_{DFDR}) \leq \alpha/2$. Consequently,

$$E_{\theta,\mathbf{y}}\left[\frac{\sum_{i=1}^{m} L_1(\psi_i, \delta_{DFDR,i}(\mathbf{y}))}{\max\{1, m - \sum_{i=1}^{m} L_2(\psi_i, \delta_{DFDR,i}(\mathbf{y}))\}}\right] \leq \frac{\alpha}{2}. \tag{4}$$

Equation (4) says that a Bayesian version of the *DFDR* is bounded by $\alpha/2$ when the Bayes rule δ_{DFDR} is used. It may be worth observing that these results apply to a very general class of multiple comparison problems. Essentially the only restriction is that the set of contrasts be finite. Indeed, these do not even have to be contrasts in the usual sense of that term. They could also, for example, be a set of independent parameters that formed a family of interest.

4. A Bayesian Version of Benjamini & Hochberg's Procedure

Next, we consider the multiple comparisons procedure proposed by Benjamini and Hochberg [1] and modified for directional testing by Williams, Jones and Tukey [13].

However, we will translate the procedure into our Bayesian framework. Define $k_{DB\,\&\,H}$ to be the largest value of $k = 1, \ldots, m$, such that

$$p_{(k)} \leq \left(\frac{\alpha}{2}\right)\left(\frac{k}{m}\right),$$

with $k_{DB\,\&\,H} = 0$ if no such value of k exists. Define $\delta_{DB\,\&\,H}(\mathbf{y}) = \mathbf{a}^{(k_{DB\,\&\,H})}$.

If $k_{DB\,\&\,H} = 0$, then the posterior expected loss for $\mathbf{a}^{(k_{DB\,\&\,H})}$ is equal to $\alpha/2$. If $k_{DB\,\&\,H} > 0$, the posterior expected loss for $\mathbf{a}^{(k_{DB\,\&\,H})}$ is given by equation (2) as

$$E_{\theta|\mathbf{y}}\left[L_{DFDR}\left(\psi, \mathbf{a}^{(k_{DB\,\&\,H})}\right)|\mathbf{y}\right] = \frac{\sum_{i=1}^{k_{DB\,\&\,H}} p_{(i)}}{k_{DB\,\&\,H}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DB\,\&\,H}}{m}\right). \tag{5}$$

From the definition of $k_{DB\,\&\,H}$, it follows that

$$p_{(i)} \leq \left(\frac{\alpha}{2}\right)\left(\frac{k_{DB\,\&\,H}}{m}\right) \text{ for } i = 1, \ldots, k_{DB\,\&\,H}. \tag{6}$$

Consequently, applying inequality (6) to equation (5), it follows that

$$E_{\theta|\mathbf{y}}\left[L_{DFDR}\left(\psi, \mathbf{a}^{(k_{DB\,\&\,H})}\right)|\mathbf{y}\right] \leq \left(\frac{\alpha}{2}\right)\left(\frac{k_{DB\,\&\,H}}{m}\right) + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DB\,\&\,H}}{m}\right) = \frac{\alpha}{2}.$$

Since this inequality holds for all $\mathbf{y}$, it implies that $r(\delta_{DB\,\&\,H}) \leq \alpha/2$, and so, just as with δ_{DFDR},

$$E_{\theta,\mathbf{y}}\left[\frac{\sum_{i=1}^{m} L_1(\psi_i, \delta_{DB\,\&\,H,i}(\mathbf{y}))}{\max\{1, m - \sum_{i=1}^{m} L_2(\psi_i, \delta_{DB\,\&\,H,i}(\mathbf{y}))\}}\right] \leq \frac{\alpha}{2}. \tag{7}$$

Equation 7 says that $\delta_{DB\,\&\,H}$ also controls our Bayesian *DFDR*.

This seems like an appropriate place to note that what has just been established (namely the fact that $\delta_{DB\,\&\,H}$ controls the *DFDR* for an arbitrary set of contrasts) is a Bayesian, rather than a sampling theory result. Indeed, Benjamini and Hochberg's [1] sampling theory procedure has only been shown to control the sampling theory *FDR* in special circumstances, such as the case of independent tests. In particular, it has not been shown to provide sampling theory control of the *FDR* when making all pairwise comparisons among a set of means in a one-way, fixed effects analysis of variance setup.

Since δ_{DFDR} is a Bayes decision rule, it must be the case that $r(\delta_{DFDR}) \leq r(\delta_{DB\,\&\,H})$. Moreover, it is also possible to show that $k_{DFDR} \geq k_{DB\,\&\,H}$ for all $\mathbf{y}$. To see this, suppose the contrary: $k_{DB\,\&\,H} = k_{DFDR} + d$ with $d > 0$. By the definition of k_{DFDR}, we must have

$$\frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{\max\{1, k_{DFDR}\}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DFDR}}{m}\right) < \frac{\sum_{i=1}^{k_{DB\,\&\,H}} p_{(i)}}{k_{DB\,\&\,H}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DB\,\&\,H}}{m}\right)$$

or

$$\begin{aligned}\frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{\max\{1, k_{DFDR}\}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DFDR}}{m}\right) &< \frac{\sum_{i=1}^{k_{DFDR}} p_{(i)} + \sum_{i=k_{DFDR}+1}^{k_{DFDR}+d} p_{(i)}}{k_{DFDR} + d} \\ &\quad + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DFDR} + d}{m}\right).\end{aligned} \tag{8}$$

Note that a strict inequality has been used here, implying that, in case of ties, k_{DFDR} would be chosen to be the largest value of k that minimizes the posterior expected loss. To continue, using the definition of $k_{DB\ \&\ H}$,

$$\sum_{i=k_{DFDR}+1}^{k_{DFDR}+d} p_{(i)} \leq d\left(\frac{\alpha}{2}\right)\left(\frac{k_{DFDR}+d}{m}\right). \tag{9}$$

Combining inequalities (8) and (9) gives

$$\frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{\max\{1, k_{DFDR}\}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DFDR}}{m}\right) < \frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{k_{DFDR}+d} + \left(\frac{\alpha}{2}\right)\left(\frac{d}{m}\right) + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DFDR}+d}{m}\right)$$

or

$$\frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{\max\{1, k_{DFDR}\}} < \frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{k_{DFDR}+d}. \tag{10}$$

If $k_{DFDR} > 0$, inequality (10) implies that $d = 0$, contrary to our initial assumption. If $k_{DFDR} = 0$, both sides of inequality (10) would be 0, contradicting the strict inequality. Thus, we have demonstrated that $k_{DFDR} \geq k_{DB\ \&\ H}$ for all $\mathbf{y}$. In other words, the Bayes rule δ_{DFDR} will always declare at least as many signs as $\delta_{DB\ \&\ H}$.

5. Simulation Results

It is important to recall that the procedure actually proposed by Benjamini and Hochberg [1] uses sampling theory p-values (one-tailed values in Williams, Jones and Tukey's [13] version), rather than posterior tail probabilities and controls the sampling theory version of the *FDR* (or *DFDR* in Willams et al.'s version). Now consider a standard multiple comparisons problem: the one-way, fixed effects analysis of variance setup, with the ψ_i chosen to be all pair-wise differences among the group means. In this case, the relevant sampling theory and Bayesian (based on a vague prior for all parameters) p-values are identical tail probabilities from the appropriate Student's t-distribution (see, for instance, Box & Tiao, [2], p. 140).

Tables 1 and 2 give the results of sampling theory simulations (based on 25,000 replications for each condition) of one-way, fixed effects ANOVA setups, considering all pair-wise differences for 10 evenly spaced means, and for 25 evenly spaced means. In these simulations, the within-group variance was set at 3.0 with $n = 3$ sample

Table 1
Sampling theory DFDR for all pair-wise comparisons made using our Bayes rule and Benjamini & Hochberg's procedure

τ : Spread of Means	10 Means		25 Means	
	δ_{DFDR}	$\delta_{DB\ \&\ H}$	δ_{DFDR}	$\delta_{DB\ \&\ H}$
0.00+	0.0204	0.0171	0.0206	0.0176
0.721	0.0062	0.0044	0.0067	0.0046
3.606	0.0005	0.0005	0.0013	0.0012
5.408	0.0001	0.0001	0.0006	0.0006
7.211	0.0000	0.0000	0.0003	0.0003
14.422	0.0000	0.0000	0.0000	0.0000

TABLE 2
Sampling theory average power for all pair-wise comparisons made using our Bayes rule and Benjamini & Hochberg's procedure

τ : Spread of Means	10 Means		25 Means	
	δ_{DFDR}	$\delta_{DB\ \&\ H}$	δ_{DFDR}	$\delta_{DB\ \&\ H}$
0.00+	0.002	0.002	0.001	0.000
0.721	0.022	0.016	0.012	0.007
3.606	0.634	0.621	0.604	0.594
5.408	0.783	0.778	0.741	0.737
7.211	0.860	0.857	0.813	0.811
14.422	0.984	0.984	0.924	0.924

Note: 25,000 replications for each condition, $n = 3$ observations per group, within degrees of freedom $\nu = 20$ for 10 means and $\nu = 50$ for 25 means, within variance $\sigma^2 = 3.0$, $\alpha/2 = 0.025$.

observations per group (so the sampling variances of the sample means are all equal to 1.0 and the within-group degrees of freedom equals 20 in the first case and 50 in the second case). In addition, we chose $\alpha = 0.05$, with the intention of controlling the sampling theory *DFDR* at $\alpha/2 = 0.025$, although we emphasize again that there is no theory to support that control for pair-wise comparisons.

The *DFDR* values in Table 1 are sampling theory averages of the sample *DFDR* used as the first term of our loss function for the two rules with two numbers of means, and a range of spacings among the means. To make the results comparable across the two setups, we used the population standard deviation (denoted here by τ) to index the spread of the means. The average power values in Table 2 are sampling theory averages of the sample per-comparison correct sign declaration rate. All four quantities for a given number of means are computed from the same 25,000 replications at each spread of the means. Note that the spread is effectively given in units equal to the standard errors of the sample means. For the spread labeled "0.00+" all population mean values were set equal, and an arbitrary ordering was chosen to evaluate the "wrong sign" errors. Both procedures conservatively control the *DFDR* for all conditions considered. The Bayes rule procedure provides slightly greater per-comparison power than that of Benjamini and Hochberg in these conditions, but the actual differences are trivial.

6. Conclusions

The decision rule δ_{DFDR} has been shown to be optimal (from a Bayesian perspective) relative to the loss function L_{DFDR} for a wide class of multiple comparison problems involving sign declarations. It has also been shown to control a Bayesian version of the directional false discovery rate (*DFDR*), as has a Bayesian version of the procedure proposed by Benjamini and Hochberg ($\delta_{DB\ \&\ H}$). There is no guarantee that δ_{DFDR} or $\delta_{DB\ \&\ H}$ will control a sampling theory *DFDR* for the case of pair-wise comparisons, although that appears to occur in the ANOVA examples given, where the two rules behave very similarly.

References

[1] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300.

[2] BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison Wesley, Reading, MA.

[3] Duncan, D. B. (1965). A Bayesian approach to multiple comparisons. *Technometrics* **7** 171–222.
[4] Jones, L. V. and Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods* **5** 411–414.
[5] Lehmann, E. L. (1950). Some principles of the theory of testing hypotheses. *Ann. Math. Statist.* **21** 1–26.
[6] Lehmann, E. L. (1957a). A theory of some multiple decision problems. I. *Ann. Math. Statist.* **28** 1–25.
[7] Lehmann, E. L. (1957b). A theory of some multiple decision problems. II. *Ann. Math. Statist.* **28** 547–572.
[8] Lewis, C. and Thayer, D. T. (2004). A loss function related to the FDR for random effects multiple comparisons. *J. Statist. Plann. Inference* **125** 49–58.
[9] Sakar, S. K. and Zhou, T. (2008). Controlling Bayes directional false discovery rate in random effects model. *J. Statist. Plann. Inference* **138** 682–693.
[10] Shaffer, J. P. (1999). A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure. *J. Statist. Plann. Inference* **82** 197–213.
[11] Shaffer, J. P. (2002). Multiplicity, directional (Type III) errors and the null hypothesis. *Psychological Methods* **7** 356–369.
[12] Waller, R. A. and Duncan, D. B. (1969). A Bayes rule for symmetric multiple comparisons problems. *J. Amer. Statist. Assoc.* **64** 1484–1503.
[13] Williams, V. S. L., Jones, L. V. and Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics* **24** 42–69.