A Bayesian Approach to the Multiple Comparisons Problem

by

Andrew A. Neath[1]
Department of Mathematics and Statistics, Southern Illinois University Edwardsville
aneath@siue.edu

Joseph E. Cavanaugh
Department of Biostatistics, The University of Iowa
joe-cavanaugh@uiowa.edu

Abstract

Consider the problem of selecting independent samples from several populations for the purpose of between-group comparisons. An important aspect of the solution is the determination of clusters where mean levels are equal, often accomplished using multiple comparisons testing. We formulate the hypothesis testing problem of determining equal-mean clusters as a model selection problem. Information from all competing models is combined through Bayesian methods in an effort to provide a more realistic accounting of uncertainty. An example illustrates how the Bayesian approach leads to a logically sound presentation of multiple comparison results.

Keywords: Bayesian information criterion, hierarchical modelling, model averaging
Running Title: Bayesian multiple comparisons

[1] Corresponding author. Address: Andrew A. Neath, Associate Professor, Department of Mathematics and Statistics, Southern Illinois University, Edwardsville, IL 62026. Phone: (618) 650-3590.

1. Introduction

Consider the problem of selecting independent samples from several populations for the purpose of between-group comparisons, either through hypothesis testing or estimation of mean differences. A companion problem is the estimation of within-group mean levels. Together, these problems form the foundation for the very common analysis of variance framework, but also describe essential aspects of stratified sampling, cluster analysis, empirical Bayes, and other settings.

Procedures for making between-group comparisons are known as multiple comparisons methods. The goal of determining which groups have equal means requires testing a collection of related hypotheses. We examine this hypothesis testing problem from a Bayesian viewpoint. In Section 2, we detail how the determination of equal mean clusters can be formulated as a Bayesian model selection problem. Posterior model probabilities are computed via the Bayesian information criterion. Bayesian model averaging is introduced as a tool for combining information from all competing models in an effort to provide a more realistic accounting of uncertainty. An example in Section 3 illustrates how the Bayesian approach to multiple comparisons testing leads to a logically sound interpretation of the results.

Section 4 addresses Bayesian estimation of within-group mean levels for the multiple comparisons problem. Similar to Stein estimation, a type of shrinkage is performed. The approach is more general, however, because shrinkage is not necessarily toward an overall mean, but rather toward means deemed likely to be equal.

2. The Multiple Comparisons Problem

Consider independent samples from I normally distributed populations with equal variances:

$$\begin{array}{c} X_{11}, X_{12}, \ldots, X_{1n_1} \sim \quad iid \quad N(\mu_1, \sigma^2) \\ \vdots \\ X_{I1}, X_{I2}, \ldots, X_{In_I} \sim \quad iid \quad N(\mu_I, \sigma^2). \end{array} \tag{2.1}$$

The goal of the multiple comparisons problem is to determine where within-group means are equal in order to create clusters of groups with equal mean levels. Thus, one is testing $H^{(a,b)} : \mu_a = \mu_b$ for each (a, b); a total of $I(I-1)/2$ distinct but related hypotheses. A typical frequentist test will decide in favor of $H^{(a,b)}$ when

$$|\bar{x}_b - \bar{x}_a| \leq Q_{a,b}.$$

The definition of $Q_{a,b}$ depends upon the approach. A point of difficulty common to classical multiple comparison testing procedures of this form is illustrated by the case where one decides in favor of $\mu_1 = \mu_2$ and in favor of $\mu_2 = \mu_3$, but against $\mu_1 = \mu_3$. Such cases are difficult to interpret since a single choice of a clustering is not obtained.

Employing a Bayesian philosophy, one might be inclined to state the goal as quantifying the evidence in favor of $H^{(a,b)} : \mu_a = \mu_b$ for each (a, b). The existence of equal mean levels is considered physically plausible for the multiple comparisons problem, so Bayesian testing of these precise hypotheses will require a measure of prior/posterior belief in $H^{(a,b)}$, and a measure of prior/posterior belief in the effect size $\delta^{(a,b)} = \mu_b - \mu_a$ if $H^{(a,b)}$ is not true. Thus, the distribution over $(\mu_1, \ldots, \mu_I)$ need consist of two components to reflect the possibility of equal means. A measure of belief in the precise nulls $\{H^{(a,b)}\}$ will be represented by point mass probabilities, while a continuous portion of the distribution will reflect belief in the size of the differences between means when $H^{(a,b)}$ is not true.

The determination of prior probabilities over the hypotheses $\{H^{(a,b)}\}$ is complicated by the fact that the collection does not consist of mutually exclusive events. For example, $H^{(1,2)}$ true $(\mu_1 = \mu_2)$ may occur with $H^{(2,3)}$ true $(\mu_2 = \mu_3)$ or with $H^{(2,3)}$ false $(\mu_2 \neq \mu_3)$. One cannot develop a prior by comparing relative beliefs in each of the hypotheses.

Furthermore, certain combinations of hypotheses in $\{H^{(a,b)}\}$ represent decisions which are logically inconsistent. For example, the event previously considered with $H^{(1,2)}$ true $(\mu_1 = \mu_2)$, $H^{(2,3)}$ true $(\mu_2 = \mu_3)$, $H^{(1,3)}$ false $(\mu_1 \neq \mu_3)$ should be assigned zero probability.

It is clear that the hypotheses must be taken as a whole when assigning prior belief. Allowable decisions can be reached through the formation of equal mean clusters among the I populations. For example, the clustering $\mu_1 = \mu_2$, $\mu_3 = \mu_4$ implies $H^{(1,2)}$ true, $H^{(3,4)}$

true, and all others false. Designating a clustering of equal mean levels will define a model nested within (2.1). When two or more means are taken as equal, we merely combine all relevant samples into one. The smaller model is of the same form as (2.1), only for $I' < I$. The problem can now be stated in terms of Bayesian model selection, where each allowable combination of hypotheses will correspond to a candidate model.

We provide a short review of Bayesian model selection in the general setting using the notation of Neath and Cavanaugh (1997). Let Y_n denote the observed data. Assume that Y_n is to be described using a model M_k selected from a set of candidate models $\{M_1, \ldots, M_L\}$. Assume that each M_k is uniquely parameterized by θ_k, an element of the parameter space $\Theta(k)$. In the multiple comparisons problem, the class of candidate models consists of all possible mean level clusterings. Each candidate model is parameterized by the mean vector $\mu = (\mu_1, \ldots, \mu_I)$ and the common variance σ^2, with the individual means restricted by the model-defined clustering of equalities. That is, each model determines a corresponding parameter space where particular means are taken as equal.

Let $L(\theta_k|Y_n)$ denote the likelihood for Y_n based on M_k. Let $\pi(k)$, $k = 1, \ldots, L$, denote a discrete prior over the models $M_1, \ldots, M_L$. Let $g(\theta_k|k)$ denote a prior on θ_k given the model M_k. Applying Bayes' Theorem, the joint posterior of M_k and θ_k can be written as

$$f(k, \theta_k|Y_n) = \frac{\pi(k)g(\theta_k|k)L(\theta_k|Y_n)}{h(Y_n)},$$

where $h(Y_n)$ denotes the marginal distribution of Y_n.

The posterior probability on M_k is given by

$$\pi(k|Y_n) = h(Y_n)^{-1}\pi(k) \int_{\Theta(k)} g(\theta_k|k)L(\theta_k|Y_n)\, d\theta_k. \tag{2.2}$$

Posterior probability on the hypothesis $H^{(a,b)}$ can be found by summing over the probabilities on those models for which $\mu_a = \mu_b$. This gives a very reasonable approach to determining the evidence in favor of each of the pairwise equalities.

We remark that placing a continuous prior in I dimensions over $(\mu_1, \ldots, \mu_I)$ will not provide a satisfactory answer to the problem of multiple comparisons testing. Under this approach, $P[H^{(a,b)}] = 0$ both *a priori* and *a posteriori*. Thus, the problem of interest is not addressed, since the precise hypotheses of primary focus are rendered impossible.

The integral in (2.2) requires numerical methods or approximation techniques for its computation. Kass and Raftery (1995) provide a discussion of the various alternatives. An attractive option is one based upon the popular Bayesian information criterion (Schwarz, 1978). Define

$$B_k = -2 \ln L(\hat{\theta}_k | Y_n) + dim(\theta_k) \ln(n),$$

where $\hat{\theta}_k$ denotes the maximum likelihood estimate obtained by maximizing $L(\theta_k|Y_n)$ over $\Theta(k)$. It can be shown under certain nonrestrictive regularity conditions (Cavanaugh and Neath, 1999) that

$$\pi(k|Y_n) \approx \frac{\exp(-B_k/2)}{\sum_{l=1}^{L} \exp(-B_l/2)}. \tag{2.3}$$

An outline of the proof is given in the Appendix.

The advantages to computing the posterior model probabilities as (2.3) include computational simplicity and a direct connection with a popular and well-studied criterion for Bayesian model selection. The justification of approximation (2.3) is asymptotic for the general case of prior $g(\theta_k|k)$, but Kass and Wasserman (1995) argue how the approximation holds under a noninformative prior on θ_k even for moderate and small sample sizes.

Now, let $\Delta = \Delta(\theta_k)$ denote a parameter of interest. For the multiple comparisons problem, focus may be on the difference between means $\delta^{(a,b)} = \mu_b - \mu_a$, or perhaps on the components of the mean vector μ. The posterior distribution of Δ given the data Y_n is

$$\begin{aligned} f(\Delta|Y_n) &= \sum_{k=1}^{L} f(k, \Delta|Y_n) \\ &= \sum_{k=1}^{L} f(\Delta|k, Y_n)\pi(k|Y_n). \end{aligned} \tag{2.4}$$

Thus, the posterior of Δ is found by taking an average of the posterior distributions under each candidate model, weighted by the posterior model probabilities. The fundamental idea behind Bayesian model averaging (BMA) is to provide a realistic accounting of the uncertainty inherent in selecting a model. Result (2.4) is the foundation of our inferential approach to the multiple comparisons problem.

3. An Example

We illustrate the Bayesian approach to multiple comparisons testing through an example involving $I = 5$ population means. The data appear in Montgomery (1997) with the objective of determining which pairs of means are significantly different using common frequentist methods of multiple comparisons. See Table 1 for the summary statistics and Figure 1 for a graphical display.

Montgomery introduced the data in the context of a completely randomized experiment designed to investigate the relationship between the tensile strength of a new synthetic fiber and the blend of cotton in the fiber. The treatment groups correspond to five different cotton blends. Five fabric specimens are tested for each blend. The response measurements reflect tensile strength (in lb/in^2). Treatments are identified in ascending order of the observed sample means.

A glance at the data suggests a potentially strong clustering of μ_1, μ_2, and a clustering to a lesser degree among μ_3, μ_4, μ_5. We shall see how these notions can be quantified from the computation of the Bayesian posterior probabilities on the pairwise equalities.

Table 1. Data for example.

group (cotton blend)	response (tensile strength in lb/in^2)	sample mean	sample s.d.
1	7,7,9,11,15	9.8	3.35
2	7,10,11,11,15	10.8	2.86
3	12,12,17,18,18	15.4	3.13
4	14,18,18,19,19	17.6	2.07
5	19,19,22,23,25	21.6	2.61

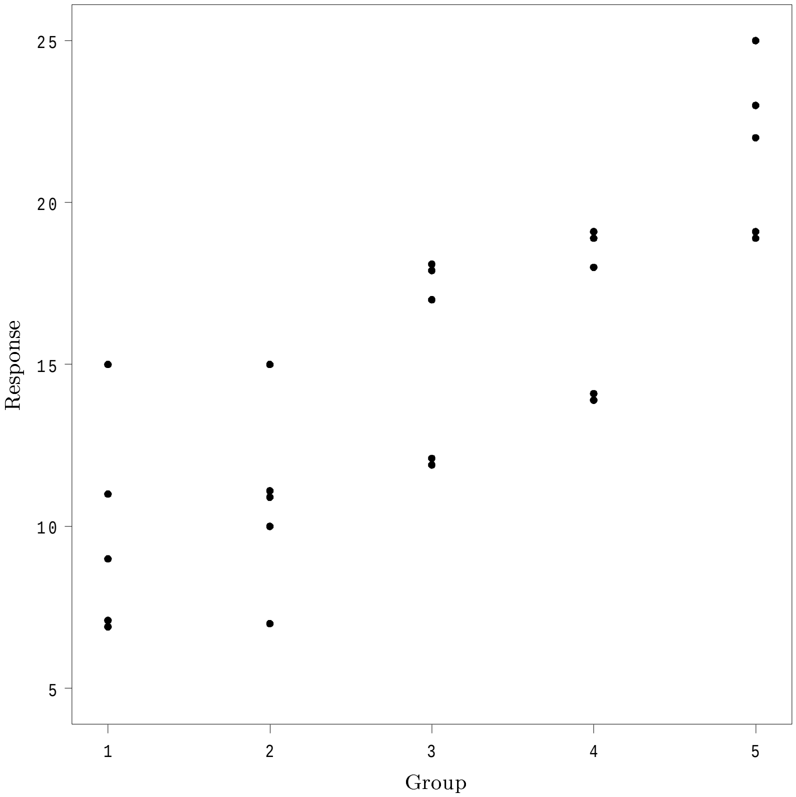

Figure 1. Scatterplot of response (tensile strength) versus group (cotton blend).

Under the setting of independent sampling with normally distributed error terms, the maximized log-likelihood is derived as

$$\ln L(\hat{\theta}_k | Y_n) = -\frac{n}{2} \ln(\hat{\sigma}^2_{(k)}) + \beta,$$

where β is a constant, $n = \sum_{i=1}^{I'} n_i$, and

$$\hat{\sigma}^2_{(k)} = \frac{1}{n} \sum_{i=1}^{I'} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2. \tag{3.1}$$

The Bayesian information for model M_k can be defined as

$$B_k = n \ln(\hat{\sigma}^2_{(k)}) + dim(M_k) \ln(n). \tag{3.2}$$

A Bayesian model selection rule would favor the model M_k which is *a posteriori* most probable, or equivalently, the model for which (3.2) is minimum.

The top five model choices and posterior probabilities are displayed in Table 2. A clear selection as "best model" is the clustering $\mu_1 = \ \mu_2, \mu_3 = \ \mu_4, \mu_5$ ($H^{(1,2)}$ true, $H^{(3,4)}$ true). It is worthwhile to note that belief in the most probable model being correct is still rather small ($\pi < .5$), so the selection of a model without a corresponding measure of uncertainty would be misleading.

Table 2. Posterior Model Probabilities.

mean clusters	$\pi(k\|Y_n)$
$\mu_1 = \mu_2, \mu_3 = \mu_4, \mu_5$	.4688
$\mu_1 = \mu_2, \mu_3, \mu_4, \mu_5$	.2286
$\mu_1, \mu_2, \mu_3 = \mu_4, \mu_5$	.1121
$\mu_1 = \mu_2, \mu_3, \mu_4 = \mu_5$	.0744
$\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$	.0554

Posterior probabilities for the top five most likely pairwise equalities are in Table 3. Posterior pairwise equality probabilities can provide a distinction that their frequentist p-value counterparts cannot. One may fail to reject a null hypothesis of equal means because either the data supports the decision of no effect or there is not enough data to detect an effect. Frequentist p-values alone are not necessarily able to distinguish between the two situations.

The hypothesis $\mu_1 = \mu_2$ is well-supported by the data ($P[H^{(1,2)}] \approx .8$), as was suspected. There is also some evidence in favor of $\mu_3 = \mu_4$ ($P[H^{(3,4)}] \approx .6$) and a non-negligible probability of $\mu_4 = \mu_5$ ($P[H^{(4,5)}] > .1$). Yet, there is good evidence against $\mu_3 = \mu_5$ ($P[H^{(3,5)}] < .02$). Let's take a closer look at the clustering among μ_3, μ_4, μ_5. Tukey's multiple comparison procedure gives a critical range of $Q = 5.37$. A pair of means is deemed not equal if the corresponding sample difference exceeds Q in magnitude. As can be seen from this example, a single clustering is not necessarily obtained. One reaches the decision of accept $\mu_3 = \mu_4$, accept $\mu_4 = \mu_5$, but reject $\mu_3 = \mu_5$. Of course, this paradoxical decision is explained by the fact that "equals" is only "no statistical significant difference," but the interpretation is still lacking any probabilistic detail. The proposed Bayesian approach bridges this gap and provides a nice presentation for multiple comparisons.

Table 3. Probabilities of Pairwise Equalities.

hypothesis	posterior
$\mu_1 = \mu_2$	.7976
$\mu_3 = \mu_4$	.6015
$\mu_4 = \mu_5$	.1200
$\mu_2 = \mu_3$	.0242
$\mu_3 = \mu_5$	.0191

To evaluate the full posterior distribution of μ, a prior must be specified under each model M_k. It is our choice to use Jeffreys' noninformative priors for $\{\mu, \sigma^2\}$ given M_k. (As mentioned earlier, the choice of a noninformative prior works well in companion to the Bayesian information approximation to model probabilities.) The components of the mean vector then have marginal posterior Student t-distributions:

$$\mu_i | M_k, Y_n \sim t\left(\hat{\mu}_{i(k)}, \frac{s_{(k)}}{\sqrt{n_i}}, n - I'_{(k)}\right), \tag{3.3}$$

where $t(m, c, v)$ represents the t-distribution with location parameter m, scale parameter c, and degrees of freedom v. The data dependent quantities within (3.3) are the sample mean $\hat{\mu}_{i(k)}$ for the cluster in model M_k containing group i, and

$$s^2_{(k)} = \frac{n}{\left(n - I'_{(k)}\right)} \hat{\sigma}^2_{(k)},$$

where $\hat{\sigma}^2_{(k)}$ is given by (3.1).

Conditional on model M_k, the posterior distribution of $\delta^{(a,b)} = \mu_b - \mu_a$ takes one of two forms. If the model restriction forces $\mu_a = \mu_b$ (that is, groups a and b are in the same cluster), then a point probability mass is placed at zero. If μ_a is allowed to differ from μ_b

(groups a and b are in different clusters), then the posterior of $\delta^{(a,b)}$, conditional on model choice, is

$$\delta^{(a,b)} \sim t\left(d_k^{(a,b)}, s_{(k)}\sqrt{\frac{1}{n_{a(k)}}+\frac{1}{n_{b(k)}}}, n-I'_{(k)}\right), \tag{3.4}$$

where $d_k^{(a,b)} = \hat{\mu}_{b(k)} - \hat{\mu}_{a(k)}$.

The BMA posterior distribution for $\delta^{(a,b)}$ is the mixture of conditional distributions defined in (2.4). A spike at zero equals the sum of probabilities over models for which $P[\delta^{(a,b)} = 0|M_k] = 1$. This sum matches with $P[\mu_a = \mu_b]$. The continuous portion is a mixture of the t-distributions in (3.4).

Figures 2 and 3 display the BMA posterior distributions for $\delta^{(1,2)}$ and $\delta^{(4,5)}$. The continuous curve is scaled to where the maximum height equals $P[\mu_a \neq \mu_b]$ so that one can make a direct comparison between the two portions.

In Figure 2, the dominant characteristic is the large spike at zero. As mentioned earlier, the data provides support for the hypothesis $\mu_1 = \mu_2$. We have $P[\delta^{(1,2)} = 0] = .7976$. In case $\mu_1 \neq \mu_2$, we believe only a small to moderate difference exists and that difference favors μ_2. One may compute probabilities backing this claim such as $P[\delta^{(1,2)} < 0] = .0561$, $P[\delta^{(1,2)} > 0] = .1463$, and $P[\delta^{(1,2)} > 5] = .0067$.

There is less belief in the precise hypothesis $\mu_4 = \mu_5$, $P[\delta^{(4,5)} = 0] = .1200$, so the continuous portion is dominant in Figure 3. One tends to believe $\mu_5 > \mu_4$, with a potentially large difference. Again, it may be informative to compute probabilities such as $P[\delta^{(4,5)} > 0] = .8736$ and $P[\delta^{(4,5)} > 5] = .3929$.

4. An Estimation Problem

We switch focus to the problem of estimating the within-group mean vector μ. The most obvious approach to this problem is to simply use the within-group sample mean vector $\hat{\mu}_w = (\bar{x}_1, \ldots, \bar{x}_I)$. At first glance, this seems like a clear choice for an estimator. However, $\hat{\mu}_w$ does not take advantage of any possible relationships between groups. In fact, $\hat{\mu}_w$ is *inadmissible* as an estimator for μ under squared error loss when $I > 3$ (Stein,

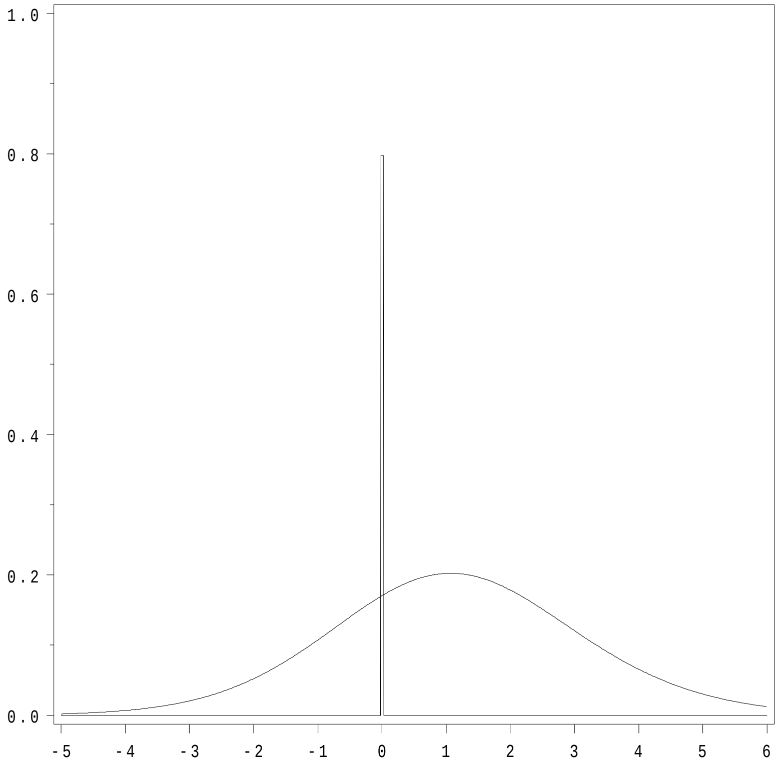

Figure 2. BMA posterior distribution for $\delta^{(1,2)}$.

1955). Improvement on $\hat{\mu}_w$ is attained using an estimator where within-group mean estimates exhibit "shrinkage" toward an overall mean.

To demonstrate the concept of shrinkage, or Stein estimation, consider the following hierarchical approach to the estimation problem at hand. We have

$$\{X_{ij}\}|\{\mu_i\}, \sigma^2 \sim \quad ind. \quad N(\mu_i, \sigma^2).$$

Suppose

$$\{\mu_i\}|\mu_o, \tau^2 \sim \quad ind. \quad N(\mu_o, \tau^2).$$

Then

$$E(\mu_i|\mu_o, \tau^2, \sigma^2, \{x_{ij}\}) = c\mu_o + (1-c)\bar{x}_i, \tag{4.1}$$

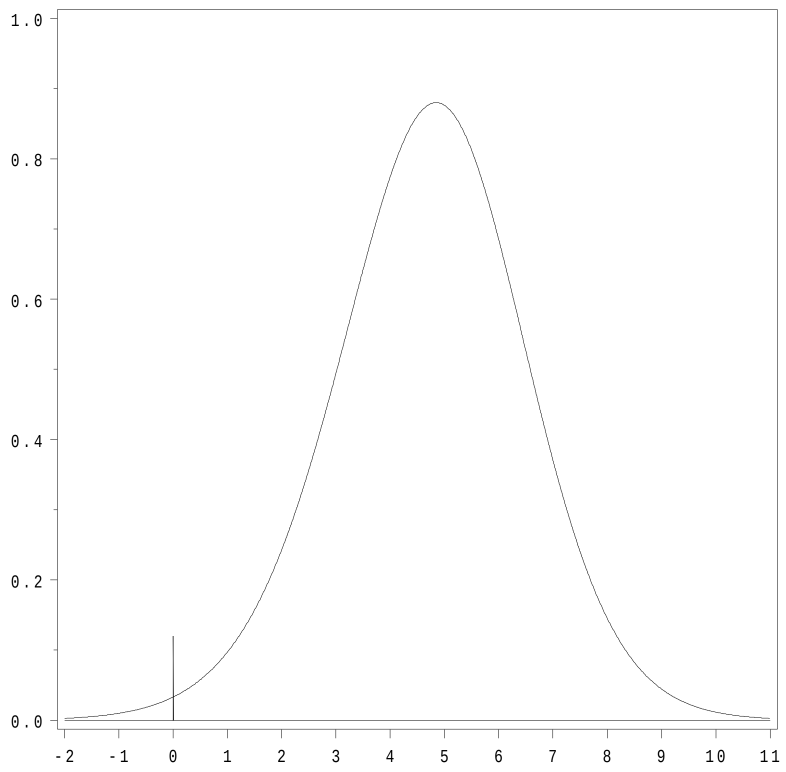

Figure 3. BMA posterior distribution for $\delta^{(4,5)}$.

where

$$c = \tau^{-2}/(\tau^{-2} + n_i\sigma^{-2}).$$

The conditional posterior mean of μ_i is a weighted average of the overall mean μ_o and the within-group sample mean $\bar{x}_i$. The weights depend upon the between-group variance τ^2 and the within-group variance σ^2/n_i. If between-group variability is small relative to within-group variability, the estimates shrink to near the overall mean μ_o. Otherwise, stronger weight is placed on individual sample means.

Expression (4.1) requires estimates of the additional parameters $(\mu_o, \tau^2, \sigma^2)$. Casella (1985) gives an empirical Bayes argument for deriving a point estimate. The estimated weights are seen as functions of the F-statistic for testing equality of all means. If F is small, the data supports the hypothesis of equal means and greater weight is placed on the overall mean. As F gets larger, weight shifts toward the individual means, consistent with the information from the data.

A Bayes empirical Bayes approach places a prior on the parameters $(\mu_o, \tau^2, \sigma^2)$. The conditional posteriors for implementing Gibbs sampling are readily attainable (Carlin and Louis, 2000). We return to the data from Table 1. Holding with our theme, a noninformative hyperprior is used. Specifically, we take

$$p(\mu_o, \ln\tau, \ln\sigma) \quad \propto \quad \tau.$$

The results are displayed in Table 4. Compare the point estimate $\hat{\mu}_{EB}$ from Table 4 to $\hat{\mu}_w$ from Table 1. The shrinkage effect is evident, but slight. The indication is against equality of *all* five mean levels.

Table 4. Interval Estimates from Bayes empirical Bayes.

parameter	posterior mean	90% interval
μ_1	10.16	(8.03,12.31)
μ_2	11.08	(8.97,13.28)
μ_3	15.39	(13.19,17.50)
μ_4	17.47	(15.29,19.59)
μ_5	21.16	(18.85,23.31)
μ_o	14.86	(9.10,20.89)
σ/n_i	0.592	(0.45,0.77)
τ	7.55	(3.14,15.55)

The idea behind our solution to the estimation problem is that intermediate models exist between the model for which all means are equal and the model for which all means are distinct. In Section 3, we showed how to compute the Bayesian probability on submodels defined by a particular clustering of equal means. An estimate of the mean vector under the BMA framework is given by

$$\hat{\mu}_{BMA} = \sum_{k=1}^{L} \hat{\mu}_{(k)} \pi(k|Y_n), \tag{4.2}$$

where $\hat{\mu}_{(k)}$ is the estimate of μ under model M_k.

Hoeting, Madigan, Raftery, and Volinsky (1999) present an overview of BMA inference with applications and optimality results. In particular, BMA is seen to improve estimation and prediction, and to adjust interval estimates which tend to be overconfident if one proceeds as if a selected model is correct with probability one.

For multiple comparisons estimation of the mean vector μ, a type of shrinkage is performed in creating the BMA estimate (4.2). However, this is a more general weighted average than a Stein estimate of the form (4.1). Shrinkage does not have to be toward an overall mean, but rather toward means deemed likely to be equal by the data.

Again consider the example. Denote the model of all means distinct as M_w and the model of all means equal as M_o. We can calculate $\pi(M_o|Y_n) < .0001$ and $\pi(M_w|Y_n) = .0554$. The posterior probability of M_w is much greater than that of M_o, which provides an alternate justification that the weight of a Stein estimate should be primarily on the individual means. Yet neither posterior probability is large, meaning neither of the models is well-supported by the data. The BMA estimate shrinks predominantly toward the most probable model

$$M_* : \quad \mu_1 = \mu_2, \mu_3 = \mu_4, \mu_5.$$

Figure 4 provides a graphical comparison of the shrinkage properties of the estimates. One can see the benefits of the BMA estimate in such an example where the data indicates several potential clusters of equal means.

The full posterior distribution on μ_i is stated by the mixture in (2.4), with distributions conditional on model choice shown in (3.3). Given a particular model choice as correct, the Bayesian intervals under the noninformative prior coincide with the standard frequentist confidence intervals. Table 5 displays both BMA intervals and the intervals under model M_*.

As a demonstration, we can focus attention on estimating μ_5. This is the treatment level where, on the basis of observed sample means, the maximum expected response occurs. The estimation problem involves the uncertainty of determining which model is correct, and the uncertainty of estimation within a given model. The choice of model M_* as correct implies

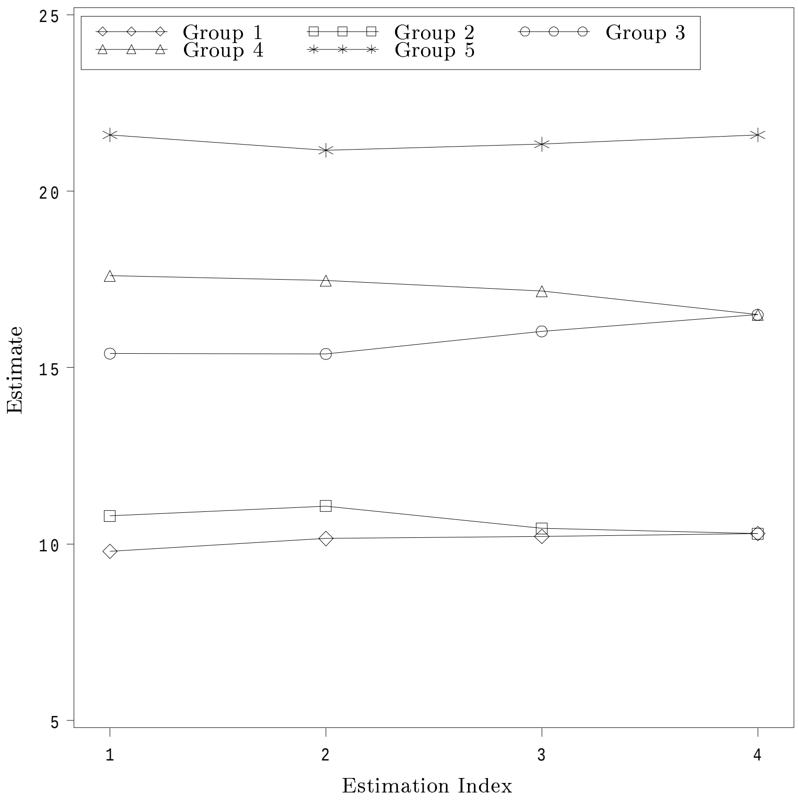

Figure 4. Group mean estimates by estimation index:
1 = model M_w; 2 = Stein; 3 = BMA; 4 = model M_*.

without question that no clustering of treatment 5 occurs. We see from Table 2, for example, that $P[\mu_4 = \mu_5] = .1200$, so the possibility of a clustering with other groups should play a role.

The BMA intervals are wider, reflecting the uncertainty associated with the selection of a clustering. Figure 5 displays the posterior distributions for μ_5 under M_* and model averaging. The greater variability under BMA for a better accounting of uncertainty is noticed. Also note the skewness of the BMA posterior due to the shrinkage property. Bayesian model averaging provides a natural approach to incorporating these desirable aspects into a solution for the estimation problem.

Table 5. Interval Estimates from BMA and M_*.

parameter	BMA 90% interval	Model M_* 90% interval
μ_1	10.22 (8.43,11.94)	10.3 (8.76,11.83)
μ_2	10.45 (8.73,12.33)	10.3 (8.76,11.83)
μ_3	16.03 (13.55,18.05)	16.5 (14.96,18.03)
μ_4	17.17 (15.13,19.94)	16.5 (14.96,18.03)
μ_5	21.33 (18.76,23.69)	21.6 (19.40,23.77)

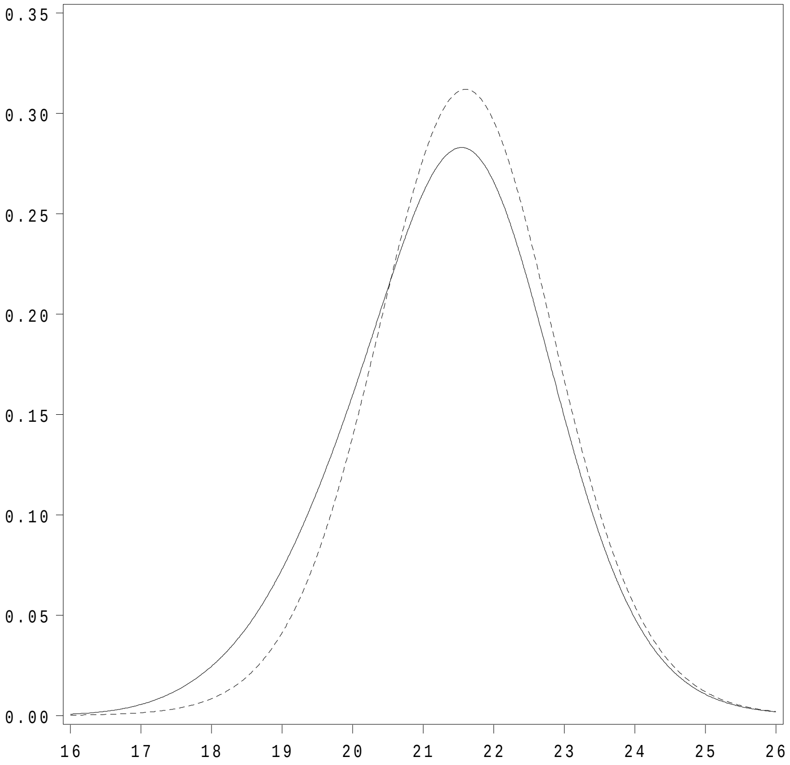

Figure 5. Posterior distributions for μ_5
under M_* (dashed curve) and model averaging (solid curve).

5. Concluding Remarks

The multiple comparisons problem is well known among statistical practitioners. Although fairly simple to state, a challenge to solving the problem lies in that one is testing several *related, precise* hypotheses. Bayesian inference has an advantage over traditional frequentist approaches to multiple comparisons testing in that degree of belief is quantified. One can avoid the illogical conclusions which arise from an "accept/reject" decision process. The Bayesian approach in this paper is novel in that the precise hypotheses used to define multiple comparisons testing are the hypotheses that are actually being tested. Bayesian approaches derived from continuous prior distributions do not possess this characteristic. We are able to compute the probability on the event of equal means, as the statement of the multiple comparisons test requires.

Acknowledgments

We wish to extend our appreciation to the referee for carefully reading the original version of our paper, and for encouraging a more extensive development of the problem area.

Appendix: Approximating the Posterior Model Probability

We present a justification for expression (2.3). A detailed proof (Cavanaugh and Neath, 1999) is rather lengthy. The purpose here is to provide a heuristic development designed to give the reader some background.

Consider a model M_k from among the candidate class $\{M_1, \ldots, M_L\}$. As expressed in (2.2), the posterior probability on M_k is given by

$$\pi(k|Y_n) = h(Y_n)^{-1}\pi(k) \int_{\Theta(k)} g(\theta_k|k)L(\theta_k|Y_n)\, d\theta_k.$$

For ease of exposition, we use a uniform prior for $\pi(k)$ (i.e., $\pi(k) = 1/L$ for all k), and a "flat," improper prior for $g(\theta_k|k)$ (i.e., $g(\theta_k|k) = 1$). (These specifications are not required for a formal proof.) We then have

$$-2\ln \pi(k|Y_n) = -2\ln \int L(\theta_k|Y_n)\, d\theta_k + c_n, \tag{A.1}$$

where c_n is constant with respect to k.

Consider the integral which appears in (A.1). To obtain an approximation to this term, we take a second-order Taylor expansion of the log-likelihood about $\hat{\theta}_k$. Since

$$\frac{\partial \ln L(\hat{\theta}_k|Y_n)}{\partial \theta_k} = 0,$$

we have

$$\ln L(\theta_k|Y_n) \approx \ln L(\hat{\theta}_k|Y_n) - \frac{n}{2}(\theta_k - \hat{\theta}_k)'[I_n(\hat{\theta}_k)](\theta_k - \hat{\theta}_k),$$

where

$$I_n(\hat{\theta}_k) = -\frac{1}{n}\frac{\partial^2 \ln L(\hat{\theta}_k|Y_n)}{\partial \theta_k\, \partial \theta_k'}$$

is the observed Fisher information matrix. Thus,

$$\int L(\theta_k|Y_n)\, d\theta_k \approx L(\hat{\theta}_k|Y_n) \int \exp\{-\frac{n}{2}(\theta_k - \hat{\theta}_k)'[I_n(\hat{\theta}_k)](\theta_k - \hat{\theta}_k)\}\, d\theta_k. \tag{A.2}$$

The integrand in (A.2) is the kernel for the multivariate normal density. Then

$$\int L(\theta_k|Y_n)\, d\theta_k \approx L(\hat{\theta}_k|Y_n)\, (2\pi)^{dim(\theta_k/2)} |nI_n(\hat{\theta}_k)|^{-1/2}. \tag{A.3}$$

We use (A.1) and (A.3) to justify writing

$$-2 \ln \pi(k|Y_n) \approx -2 \ln L(\hat{\theta}_k|Y_n) + dim(\theta_k) \ln(n) + \beta_n,$$

where β_n represents those terms that are either constant with respect to k or bounded as the sample size grows to infinity. Define

$$B_k = -2 \ln L(\hat{\theta}_k|Y_n) + dim(\theta_k) \ln(n).$$

Then

$$\pi(k|Y_n) \approx \exp(-B_k/2) \exp(-\beta_n/2).$$

With respect to the candidate model class $\{M_1, \ldots, M_L\}$, we obtain

$$\pi(k|Y_n) \approx \frac{\exp(-B_k/2)}{\sum_{l=1}^{L} \exp(-B_l/2)}.$$

References

Carlin, B. and Louis, T. (2000). Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall, New York.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, 39, 83-87.

Cavanaugh, J. and Neath, A. (1999). Generalizing the derivation of the Schwarz information criterion. *Communications in Statistics*, 28, 49-66.

Hoeting, J., Madigan, D., Raftery, A. and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14, 382-401.

Kass, R. and Raftery, A. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90, 773-795.

Kass, R. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90, 928-934.

Neath, A. and Cavanaugh, J. (1997). Regression and time series model selection using variants of the Schwarz information criterion. *Communications in Statistics*, 26, 559-580.

Montgomery, D. (1997). Design and Analysis of Experiments. Wiley, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium*, 197-206. University of California Press.