Multiple Comparisons, Multiple Tests, and Data Dredging: A Bayesian Perspective

by

Donald A. Berry*
University of Minnesota
Technical Report No. 486
April, 1987

* Research supported in part by the National Science Foundation under grant no. DMS 85-05023. Invited paper to the Third Valencia International Meeting on Bayesian Statistics, 1-5 June 1987.

1. Introduction

An experiment compares a new procedure with a control on the basis of k variables. The results suggest that new is better than control for some variables but not for others. Statistical hypothesis tests find that new is "significantly better" ($\alpha = .05$) for two of the variables.

Classical analysis does not stop here. Suppose new and control are actually identical insofar as these k variables are concerned; this is the intersection of the k individual null hypotheses. Then one expects new to be statistically significantly better for 5% of the variables. For example, if $k = 40$ then observing two significant results is actually expected when all 40 null hypotheses are true.

Suppose that the k variables $X_1, \ldots, X_k$ are differences, new minus control, standardized to have unit variance. Suppose further, that each X_i is normally distributed with mean μ_i and that the X_i's are independent given the μ_i's. The ith null hypothesis is $\mu_i = 0$. The intersection of all k null hypotheses is H_0: $\mu_1 = \ldots = \mu_k = 0$. The null probability of $X_i > 1.645$ is .05 since $\Phi^{-1}(.95) = 1.645$, the 95th percentile of the standard normal. But the null probability that at least one $X_i > 1.645$ is $1-.95^k$. This obviously increases rapidly; for $k = 40$ it is about .87. So the null probability of at least one rejection can be much larger than the "nominal level" .05. Classical statistical wisdom dictates an adjustment. One way to make the "actual level" equal to .05 is to increase the rejection limits for each X_i from 1.645 to $\Phi^{-1}(.95^{1/k})$. For example, $.95^{1/40} = .9987$ and $\Phi^{-1}(.9987) = 3.00$. So to attain significance for the ith variable <u>at the .05 level</u> requires

$X_i > 3.00$.

This seems ludicrous from a scientific point of view: How can the mere fact that bilirubin levels were measured (not what they were, just that they were measured!) affect inferences about blood pressure? An investigator who carried out only one test might find a significant difference whereas the same difference would not have been significant had the investigator tested enough other variables. (There is a strong temptation to cheat--cf. Berger and Berry (1987). Especially since such cheating is impossible to uncover--it depends on the intentions of the investigator rather than on the data, which are unadulterated. And it would be regarded as cheating at most by those few scientists who understand this statistical construction.)

Such classical statistical adjustments also seem inconsistent with a Bayesian point of view; the mere fact that variables were measured is irrelevant to the current distribution of the μ_i's. Of course, the actual observations (or partial information about these observations) can change the current distribution. The new distribution is a function of those observations and so is random--its average being the current distribution.

Still, this issue is not clearcut. For example, when observing a large number of exchangeable random variables, some will be larger than others. Selecting those that are extreme can be misleading. In both this example and the previous setting, the infinitely careful Bayesian need not worry, but the Bayesian who assumes a prior distribution without careful reflection may obtain a posterior distribution that is far from what it should be. While this statement is true generally, the issue is more critical when making multiple

inferences than when making inferences in a univariate setting.

Adjustments for simultaneously testing many null hypotheses are similar to those that arise in many guises in classical statistics. Historically the most important of these has been the problem of "multiple comparisons," in which many treatments (and their interactions) are compared on the basis of measuring some variable for each treatment (see Section 3). Another is applying several different statistical tests of the same hypothesis to the same data (t-test, signed-rank test, etc.). An area which has recently been the focus of much research is "interim analysis," in which tests of an hypothesis are carried out periodically as data accumulate.

The Bayesian approach to interim analysis is straightforward (Berry 1985, 1987) and uncomplicated by some of the considerations of multiple comparisons and testing many variables (see Section 4). The distinction is that in the problem of interim analysis (and also that of several different tests for the same hypothesis) the hypothesis being tested involves a one-dimensional parameter (e.g., $\mu_1 = 0$ as opposed to $\mu_1 = \ldots = \mu_k = 0$). This might seem like a trivial distinction--one simply applies Bayes's theorem in both cases. But, much greater care seems necessary in assessing prior information when the null hypothesis is more complicated.

Still another area of some concern to classical statistics is variable selection in regression. Letting the data indicate which of many variables to use in regression can greatly exaggerate the appropriateness of such a data-indicated model from a classical statistical point of view. However, there would be no objection to the same model if the variables had been selected in

advance, or separate from the data! Suppose a researcher uses a linear model involving a number of independent variables. If the researcher chose the variables in advance then the results are believable; otherwise, they are not. If you cannot tell which of the two from a report of the study, then (as I have heard classical statisticians advise) you have to write to the researcher to find out which! (What to do if the researcher has since died? Randomize?) It is hard for me to understand how classical statisticians can continue to adhere to a philosophy that takes them down so many roads that lead to nonsense.

A more general question of great importance to statistics and all of science is whether and how it is possible to learn from data. Sounds silly! Statistics <u>is</u> learning from data. But classical statistics qualifies this statement. A fundamental tenet of classical statistics is that one cannot test a hypothesis using the same data that generated the hypothesis. If one notices a tendency through "data dredging" then one must get <u>another</u> data set to verify that the tendency is real. The problem, classical statisticians say, is that there are so many tendencies that could be dredged up, noticing one tendency is not very surprising.

A Bayesian can calculate the posterior probability that any given tendency is real, but requires a prior probability. It seems impossible to, in advance of an experiment, assess one's prior probability for each tendency that might arise. But is it possible to assess one's <u>prior</u> probability <u>after</u> seeing the data? I think it is possible, but it is very difficult. Obviously, it is wrong to incorporate the same data into a posterior probability twice. So one must be able to say with some confidence, "This is what I thought before."

Alternatively, one can simply assess one's current probability after having seen the data, eschewing the use of Bayes's theorem and relying instead on one's internal analogue. We do this all the time. Our internal Bayes's theorem may not process information as well as does the real thing, and some of us may process it incredibly badly, but we are not likely to use the data twice. To check an assessor's ability in this regard, apply Bayes's theorem in reverse, dividing the posterior by the likelihood to yield the prior. This may not be possible and, if possible, when there are zeroes in the likelihood the result is not uniquely defined. If it is not possible then the assessor can be instructed on removing this inconsistency. In any case, it is appropriate to adjust the posterior if the assessor "never could have had that prior."

Still a third possibility is to find people who haven't seen the data and have them serve as surrogate assessors. This procedure can serve to educate the assessor, and can be combined with the other two procedures.

In the next section I discuss the possibility of, and difficulties associated with, making Bayesian inferences concerning hypotheses generated by the data. In Section 3 I define multiple comparisons and multiple tests. Section 4 draws a parallel with the so-called empirical Bayes problem and suggests this as a way to view some problems in multiple inference.

My goal throughout is not to give results which are immediately useful to the practitioner, but only to elucidate the major issues. So the examples and settings I use are rather simple.

2. Data Dredging; Simultaneous Learning and Testing

Consider this scenario. A large study was conducted by randomly selecting 30-year-old men and following them for a long period of time. The investigators measured hundreds of variables at five-year intervals, with no particular plan for testing them all. Some conclusions are boringly predictable: men who were overweight tended to die at an earlier age and were generally less healthy; similarly for men who smoked cigarettes. But there's something new and quite unexpected: men who chewed gum regularly lived six years longer on average than men who never chewed gum! This difference is "highly statistically significant": nominal P-value < .01. Moreover, this difference persists upon adjusting for all available covariates.

As I have indicated, classical statisticians would consider the number of tests that had been carried out and suitably adjust the P-value upwards, for example, using Bonferroni's inequality. In particular, if the number of tests is sufficiently large, the result will no longer be significant. And if this number is not available then a good classical statistician would say that correct inferences regarding this issue require another study.

What about the Bayesian point of view? Figure 1 shows the likelihood function for μ, the mean increment in length of life (in years) as a result of chewing gum. This is reasonably approximated by $N(6, 2^2)$. In particular, the likelihood ratio of $\mu = 6$ (obviously the extreme case) compared with $\mu = 0$ is about 89.

My prior distribution on μ (which I can assess unencumbered by knowing the data because I also know that the data are fake!) is approximately

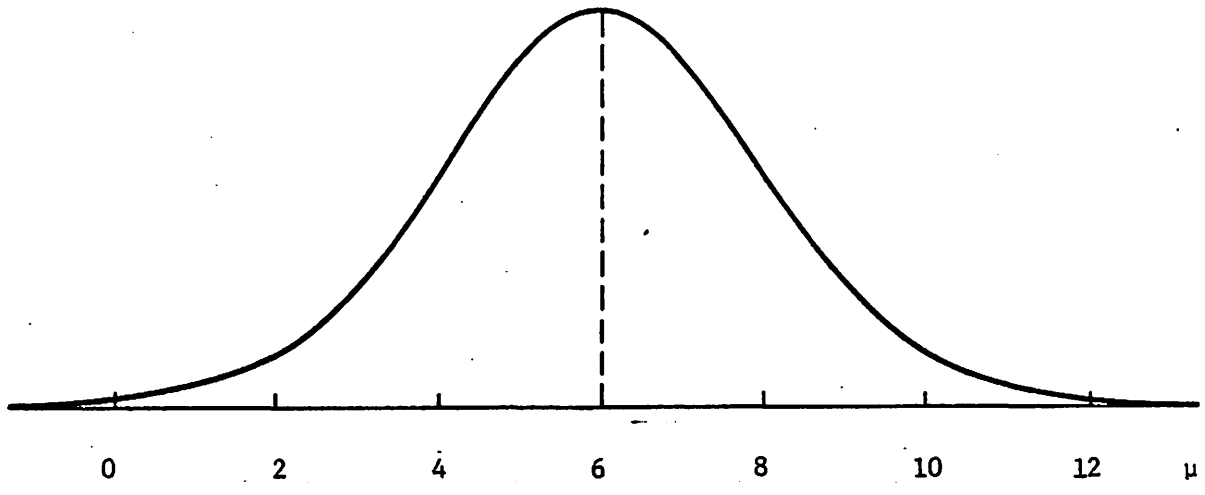

Figure 1. Likelihood of μ.

$.6\delta_0 + .4N(-.1, 1.2^2)$--the probability of $\mu = 0$ is .6 and the rest of my probability is dispersed rather near 0. So I'm not convinced that chewing gum has no effect but I doubt that it has much. (The negative mean, $E(\mu) = -.04$ years, reflects my pessimism regarding the healthfulness of a regular intake of sugar. This pessimism is partially balanced by the possibility that the miniscule amount of exercise one gets while chewing gum might be beneficial!)

To be somewhat more general, suppose that the prior distribution of μ is

$$p\delta_0 + (1-p)N(\mu_0, \sigma_0^2). \qquad (1)$$

Then the posterior distribution when the likelihood is $N(6, 2^2)$ is

$$(\mu|\text{data}) \sim p'\delta_0 + (1-p')N(\mu_1, \sigma_1^2) \qquad (2)$$

where

$$\sigma_1^2 = 4\sigma_0^2/(4 + \sigma_0^2),$$

$$\mu_1 = \sigma_1^2[6/4 + \mu_0/\sigma_0^2],$$

and p' is defined by the posterior odds ratio:

$$\frac{p'}{1-p'} = \frac{p}{1-p} \cdot \frac{\sigma_0 \exp\{-\frac{1}{2} \cdot \frac{36}{4}\}}{\sigma_1 \exp\{-\frac{1}{2}[\frac{36}{4} + \mu_0^2/\sigma_0^2 - \mu_1^2/\sigma_1^2]\}}$$

The mean of this distribution is $E(\mu|\text{data}) = (1-p')\mu_1$.

Substituting $\sigma_0 = 1.2$, $\mu_0 = -.1$, and $p = .6$ gives

$$(\mu|\text{data}) \sim .37\delta_0 + .63N(1.51, 1.03^2),$$

which has mean $E(\mu|\text{data}) = .95$ years. So the data has little effect on my rather strongly held opinion that chewing gum cannot increase one's life expectancy by anything like six years.

I will consider two alternative prior distributions. The first Bayesian is rather cavalier, and the second's prior has been affected by the data. These types of behavior are never good, but I want to show how bad they can be in the

current setting. Though both priors are rather extreme to best make my point, I have kept the prior probability of $\mu = 0$ at $p = .6$ to facilitate comparison.

The Bayesian who goes overboard in being open-minded might choose $p = .6$, $\mu_0 = 0$, and $\sigma_0^2 = 10^2$ in (1). Substituting these into (2) gives

$$(\mu|\text{data}) \sim .09\delta_0 + .91N(5.77, 1.96^2),$$

which has mean $E(\mu|\text{data}) = 5.2$ years. Obviously, the data have substantially changed this prior. (Incidentally, the Bayesian who assumes $\sigma_0^2 \longrightarrow \infty$ in (1) is actually being dogmatic rather than open-minded. For, $p' \longrightarrow 1$ for any μ_0 and $p > 0$; and so the distribution of $(\mu|\text{data}) \longrightarrow \delta_0$! Also, of course, the distribution of $(\mu|\text{data}) \longrightarrow \delta_0$ as $\sigma_0^2 \longrightarrow 0$, though the functional relation above gives $p' \longrightarrow p$. (Neither of these results depends on the data.) Between these two extremes, p' has a minimum value which for these data and $\mu_0 = 0$ occurs at $\sigma_0^2 \doteq 5.7^2$; the minimum value of p' when $p = .6$ is .076, which is rather close to $p' = .09$ corresponding to $\sigma_0^2 = 10^2$.)

It is not an easy matter to put Figure 1 out of one's mind having seen it, or to ever be convinced of having put it out of mind! The Bayesian who forms an opinion after looking at the data and then updates via Bayes's theorem using <u>the same data</u> is obviously acting unreasonably. As an extreme case, suppose $p = .6$ and the prior mean and variance of $(\mu|\mu \neq 0)$ are the maximum likehood estimates: $\mu_0 = 6$ and $\sigma_0^2 = 2^2$. Substituting these into (2) gives

$$(\mu|\text{data}) \sim .02\delta_0 + .98N(6, 2),$$

which has mean $E(\mu|\text{data}) = 5.9$ years. The prior weight of $p = .6$ on $\mu = 0$ has been essentially annihilated; such a Bayesian is pretty convinced that μ is near 6 years. The data have virtually the entire say when allowing one aspect of the prior to depend on the data.

I hope this example convinces you first that it is possible to test hypotheses using the data that generated them, and second that doing so is not without peril.

3. Multiple Comparisons and Multiple Tests

To set up some terminology and conventions for the next section, consider k populations with means $\mu_1, \ldots, \mu_k$. Observations $X_1, \ldots, X_k$ are available on these populations such that $EX_i = \mu_i$. Given $\mu_1, \ldots, \mu_k$ the X_i's are independent. Let $X_{(1)} \leq \ldots \leq X_{(k)}$ be the ordered observations and $\mu_{(i)} = EX_{(i)}$ (so $\mu_{(i)}$ is the mean of the population with the ith smallest observation--the $\mu_{(i)}$'s may not be ordered).

The following two problems involve multiple inferences:

Multiple Comparisons: Based on $X_1, \ldots, X_k$, make inferences concerning the relationships among the various μ_i's.

Multiple Tests: Based on $X_1, \ldots, X_k$, make inferences concerning the various μ_i's individually.

An example of the first is testing the hypothesis that $\mu_{(1)} = \mu_{(2)}$. An example of the second is testing the hypothesis that $\mu_{(1)} = 0$.

There are at least two types of multiple testing problems:

(i) The populations are k different treatments and the X_i's refer to the same measurement (blood pressure, e.g.), and

(ii) There is one treatment or experimental setting and the populations refer to k different measurements (blood pressure, heart rate, bilirubin levels, etc.).

Obviously, in a multiple comparisons problem, only type (i) is appropriate: one would hardly be interested in the difference between mean blood pressure and mean heart rate. Some of the discussion so far in this article deals with (ii) in the context of multiple tests. The next section focuses mainly on (i) in the context of multiple comparisons and multiple tests.

4. An Empirical Bayes Connection

Though the setup in this section is rather simple, the conclusions correspond rather closely with the way Bayesians should think about comparisons and tests when there are multiple treatments. For example, to make inferences about $\mu_{(1)}$ one may need to know all the data, not just $X_{(1)}$. Also $\mu_{(1)}$ is typically positively correlated with $X_{(i)}$, $i = 1, \ldots, k$.

Suppose that the treatment responses are

$$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, k.$$

The μ_i's are unknown and so are themselves random variables, their joint

distribution reflecting information about, and relationships among, the various treatment responses.

While most of this section treats problems of type (i) defined in the previous section, I want to consider type (ii) problems briefly. Suppose the X_i's are measurements on k different variables in the same experimental setting. It seems reasonable a priori to consider the possibility that the μ_i are independent. But one might also allow for the possibility that they are related. For example, a drug that decreases blood pressure is likely to affect (positively or negatively) heart rate, left ventricular ejection fraction, vascular resistance, etc. Such possibilities should be considered in multiple testing problems. It may be wrong to suppose independence but it would also be wrong to suppose that the μ_i's are positively correlated, say.

If the μ_i's are independent a priori then they will also be independent given $X_1, \ldots, X_k$. So in this case the posterior distribution of μ_i given $X_1, \ldots, X_k$ is simply the posterior distribution of μ_i given X_i, and making inferences about μ_i is a one-dimensional problem. In particular, <u>there is no multiple inference issue</u>: the distribution of $\mu_{(i)}$ depends on $X_1, \ldots, X_k$ only through $X_{(i)}$.

While it may be reasonable for someone to regard the means of k <u>variables</u> (type (ii)) as being independent, this seems less reasonable for k <u>treatments</u> (type (i)). Independence across treatments assumes very firm information about the treatments and the experimental setting. The careful probability assessor may well recognize that there is an underlying unknown effect which influences all observations similarly, irrespective of treatment. For example, consider a

clinical trial involving several treatments for breast cancer. This disease continues to be diagnosed earlier and earlier in its cycle. Thus every treatment should be more effective now than it ever was before, though how much more effective would not be clear in advance. This is certainly true of no treatment because a patient "treated" earlier will obviously live longer after such "treatment." In addition, the entrance criteria and the way these criteria are administered by clinicians vary from one trial to the next; the differences can seem minor but still show up in the results very dramatically. Because individual trials tend to involve rather homogeneous populations, the various treatments used in the current trial may seem more like each other than a single treatment seems like itself in previous trials. (A consequence of these considerations is that there is no such thing as a "known" treatment, one whose effectiveness in an experiment can be predicted up to statistical error. Even if an experimenter is meticulous about ensuring that all aspects of the current trial duplicate those of a previous trial, time and its many covariates will be different.)

There are many ways that the μ_i can be dependent. One suggested by the previous paragraph is that $(\mu_1, \ldots, \mu_k)$ is a random sample from some distribution G, which is itself unknown. This is precisely the setup assumed in the "empirical Bayes" problem proposed by Robbins (1956). Converting one's available information about G into a probability distribution gives rise to what Deely and Lindley (1981) call a "Bayes empirical Bayes" problem; see also (Berry and Christensen 1979). When I say empirical Bayes I mean Bayes empirical Bayes. The empirical Bayes objective is usually to estimate G or the various

μ_i. The adaptation here is to hypothesis tests concerning the μ_i.

To keep things reasonably simple, suppose treatment i has either has no effect ($\mu_i = 0$) or a known positive effect ($\mu_i = 1$). Distribution G has parameter p, which is the probability that any particular treatment has no effect:

$$(\mu_i|p) \sim p\delta_0 + (1-p)\delta_1.$$

Conditional on p, the μ_i are independent, and so they are exchangeable. (This assumption is clearly inappropriate in type (ii) problems.) The proportion p of treatments with no effect is unknown, with a priori density uniform on (0,1): $\pi(p) = 1$, which then implies the initial distribution for G and for the μ_i. In particular, the marginal distribution of μ_i is

$$\mu_i \sim \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1.$$

Consider X_1, the observed response to treatment 1. Its conditional distribution given p is

$$(X_1|p) \sim pN(0, 1) + (1-p)N(1, 1);$$

unconditionally,

$$X_1 \sim \frac{1}{2}N(0, 1) + \frac{1}{2}N(1, 1).$$

The density of p on (0,1) given X_1 is

$$\pi(p|X_1) = 2\,\frac{p\cdot\exp[-\frac{1}{2}X_1^2] + (1-p)\cdot\exp[-\frac{1}{2}(X_1-1)^2]}{\exp[-\frac{1}{2}X_1^2] + \exp[-\frac{1}{2}(X_1-1)^2]}$$

$$= 2\,\frac{p + (1-p)\exp[X_1-1/2]}{1 + \exp[X_1-1/2]}$$

If, for example, $X_1 = 1/2$, halfway between the two candidates for μ_1, then $\pi(p|X_1=1/2) = \pi(p)$, corresponding with the obvious fact that $X_1 = 1/2$ is noninformative as regards $\mu_1 = 0$ vs $\mu_1 = 1$. Also, $\pi(p|X_1) \longrightarrow 2p$ or $2(1-p)$ according as $X_1 \longrightarrow -\infty$ or $X_1 \longrightarrow +\infty$, equivalent to actually observing $\mu_1 = 0$ and $\mu_1 = 1$, respectively.

Now consider the distribution of μ_1 given X_1. The new probability of $\mu_1 = 0$ is

$$P(\mu_1=0|X_1) = \int_0^1 \frac{2pdp}{1 + \exp(X_1-1/2)} = \frac{1}{1 + \exp(X_1-1/2)}, \qquad (3)$$

a decreasing function of X_1. So we have

$$(\mu_1|X_1) \sim \frac{1}{1 + \exp(X_1-1/2)}\,\delta_0 + \frac{\exp(X_1-1/2)}{1 + \exp(X_1-1/2)}\,\delta_1.$$

Again, $X_1 = 1/2$ leaves the prior unchanged.

Now consider responses X_1 and X_2 on treatments 1 and 2. We have

$$(X_1, X_2 | p) \sim [pN(0, 1) + (1-p)N(1, 1)]^2$$

and so

$$\pi(p|X_1, X_2) \propto \prod_{i=1}^{2} [p + (1-p)\exp(X_i - 1/2)].$$

It follows that

$$\begin{aligned} P(\mu_1=\mu_2=0|X_1, X_2) &\propto 2 \\ P(\mu_1=0, \mu_2=1|X_1, X_2) &\propto \exp(X_2 - 1/2) \\ P(\mu_1=1, \mu_2=0|X_1, X_2) &\propto \exp(X_1 - 1/2) \\ P(\mu_1=\mu_2=1|X_1, X_2) &\propto 2\cdot\exp(X_1+X_2-1). \end{aligned}$$

As a consequence,

$$P(\mu_1=0|X_1, X_2) = \frac{2 + \exp(X_2 - 1/2)}{2 + \exp(X_2 - 1/2) + \exp(X_1 - 1/2) + 2\cdot\exp(X_1+X_2-1)}. \qquad (4)$$

This tends to 1 or 0 according as $X_1 \longrightarrow -\infty$ or $+\infty$. And it tends to $2/[2 + \exp(X_1 - 1/2)]$ or $1/[1 + 2\cdot\exp(X_1 - 1/2)]$ according as $X_2 \longrightarrow -\infty$ or $+\infty$. These latter conclusions are the same as the conditional distribution of μ_1

given X_1, having already conditioned on μ_2 (the conditional distribution of μ_1 given μ_2 being

$$(\mu_1|\mu_2) \sim \begin{cases} \frac{2}{3}\delta_0 + \frac{1}{3}\delta_1 & \text{if } \mu_2 = 0 \\ \frac{1}{3}\delta_0 + \frac{2}{3}\delta_1 & \text{if } \mu_2 = 1). \end{cases}$$

Compare (3) with (4). Obviously, they are equal if $X_2 = 1/2$--there is no information about p in $X_2 = 1/2$. When X_2 is greater than 1/2 then (4) > (3): if treatment 2 gives a large response then it is more difficult to decide that treatment 1's effect is small. For example, if $X_2 = 2$ then .43 has to be subtracted from X_1 to keep the same probability of $\mu_1 = 0$. On the other hand, if $X_2 = -2$ then to keep the probability of $\mu_1 = 0$ the same requires that .58 be added to X_1.

Consider the impact that this has on the question of multiple tests. A researcher reports that $X_{(1)} = -1$. If k = 1 then the probability of $\mu_{(1)} = 0$ from (3) is $1/(1 + \exp(-3/2)) = .82$. But if k = 2 then the probability of $\mu_{(1)} = 0$ from (4) is

$$P(\mu_{(1)}=0|X_{(1)}=-1,X_{(2)}) = \frac{2 + \exp(X_{(2)}-1/2)}{2 + \exp(X_{(2)}-1/2) + \exp(-3/2) + 2\cdot\exp(X_{(2)}-2)}.$$

This is shown in Table 1 with, for comparison, the joint probabilities of $\mu_{(1)}$ and $\mu_{(2)}$. So if $X_{(1)}$ is much less than $X_{(2)}$ then the evidence in favor of $\mu_{(1)} = 0$ is not as strong as if $X_{(1)}$ and $X_{(2)}$ were both small.

TABLE 1

Joint distribution of $(\mu_{(1)},\mu_{(2)})$ given $X_{(1)} = -1$ as a function of $X_{(2)}$

$X_{(2)}$	-1	0	.5	1	2	3	∞
$P(\mu_{(1)}=0 \mid X_{(1)}=-1, X_{(2)})$	.87	.84	.82	.79	.74	.71	.69
$P(\mu_{(2)}=0 \mid X_{(1)}=-1, X_{(2)})$	.87	.72	.61	.48	.26	.11	0
$P(\mu_{(1)}=\mu_{(2)}=0 \mid X_{(1)}=-1, X_{(2)})$	.79	.65	.55	.43	.23	.10	0
$P(\mu_{(1)}=0, \mu_{(2)}= \mid X_{(1)}=-1, X_{(2)})$	.09	.20	.27	.36	.51	.61	.69
$P(\mu_{(1)}=1, \mu_{(2)}=0 \mid X_{(1)}=-1, X_{(2)})$	.09	.07	.06	.05	.03	.01	0
$P(\mu_{(1)}=\mu_{(2)}=1 \mid X_{(1)}=-1, X_{(2)})$	.04	.09	.12	.16	.23	.27	.31
$E(\mu_{(2)}-\mu_{(1)} \mid X_{(1)}=-1, X_{(2)})$	0	.12	.21	.31	.49	.60	.69
$E(\mu_{(2)} \mid X_{(2)})-E(\mu_{(1)} \mid X_{(1)}=-1)$	0	.20	.32	.44	.64	.74	.82

The joint probabilities of $\mu_{(1)}$ and $\mu_{(2)}$ are especially relevant for the question of multiple comparisons. For example, the penultimate row of Table 1 shows how much less $\mu_{(1)}$ is expected to be than $\mu_{(2)}$ as a function of $X_{(2)}$ when $X_{(1)} = -1$. The last row of the table shows the corresponding difference ignoring the relationship between μ_1 and μ_2. These two rows are analogous respectively, to the classical statistician considering and not considering the multiple comparisons question. Obviously, ignoring the relationship exaggerates the difference between $\mu_{(1)}$ and $\mu_{(2)}$.

The more general case of making inferences about the μ_i's given $X_1, \ldots, X_k$ involves more complicated calculations, but no new ideas. For larger k any inference about μ_1 (or $\mu_{(1)}$) depends less on X_1 (or $X_{(1)}$) and more on responses to the other treatments. For example,

$$P(\mu_1=0|X_1,\ X_2=\ldots=X_k \longrightarrow -\infty) = \frac{k}{k + \exp(X_1-1/2)}$$

and

$$P(\mu_1=0|X_1,\ X_2=\ldots=X_k \longrightarrow +\infty) = \frac{1}{1 + k\cdot\exp(X_1-1/2)}.$$

So it is easier to conclude that μ_1 is small when the responses to the other treatments are small but it is difficult (though not impossible!) to conclude that μ_1 is small when the other responses are large. In particular, given X_1 and $X_2 = \ldots = X_k \longrightarrow +\infty$, for the probability of $\mu_1 = 0$ to be greater than 1/2 requires $X_1 < 1/2 - \log k$.

An empirical Bayes approach which is quite promising for the multiple comparisons problem uses mixtures of Dirichlet processes (Antoniak 1974, Berry and Christensen 1979). In this approach, estimating G and the various μ_i requires finding the posterior probabilities of all possible combinations of equality and inequality among the μ_i. For example, when k = 3 there are five possibilities: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_1$, $\mu_1 = \mu_2 \neq \mu_3$, $\mu_1 = \mu_3 \neq \mu_2$, $\mu_1 \neq \mu_2 = \mu_3$, and $\mu_1 = \mu_2 = \mu_3$. One then calculates $P(\mu_1 = \mu_2|X_1,X_2,X_3)$, say, by adding the second and fifth of these. A less than attractive aspect of using mixtures of Dirichlet processes is that the number of terms in the mixture increases very

fast as a function of k (Berry and Christensen 1979).

The empirical Bayes approach in settings of multiple comparisions and multiple tests gives results which incline toward the classical statistical view. For example, $X_{(1)}$ and $X_{(2)}$ can be further apart than would be expected by the naive analyst and still be consistent with the null hypothesis $\mu_{(1)} = \mu_{(2)}$. And a researcher cannot simply report $X_{(1)}$ as an estimate of $\mu_{(1)}$ or as a statistic for tests concerning $\mu_{(1)}$; the rest of the data $(X_{(2)}, \ldots, X_{(k)})$ also contains relevant information. But while classical "adjustments" depend only on k, empirical Bayes adjustments depend on the actual data. In analogy with classical adjustments, there would also be a Bayesian adjustment to inferences concerning $\mu_{(1)}$ if for some reason the Bayesian does not know all the data but only knows $X_{(1)}$ and k. I have not addressed this problem here.

5. Conclusions

From a Bayesian point of view it is possible to dredge data to generate hypotheses and then to test these hypotheses using the same data. However, the path to such inferences is perilous.

Correcting for comparisons and tests involving multiple treatments, so widely espoused by classical statisticians, has an analogue in Bayesian statistics. Namely, one assumes that the treatments are themselves sampled from an unknown distribution, as in the empirical Bayes problem. When the k treatments have means μ_i that are exchangeable a priori, the following rough

interpretation is consistent with the empirical Bayes approach, and it seems reasonable in the problem at hand. The posterior distribution of μ_i is pulled toward treatment i response X_i, but also in the directions of the responses to the other treatments. For any pair of treatments (i,j), the estimated distance between μ_i and μ_j given X_1, ..., X_k is less than that between X_i and X_j, but the signs of the two differences are the same. In particular, if $X_i = X_j$ then μ_i and μ_j are exchangeable a posteriori as well as a priori.

The empirical Bayes approach is more relevant for simultaneous inferences concerning many treatments, problem (i) of Section 3, than concerning many variables (same "treatment"), problem (ii) of Section 3. Regarding the latter, I do not see that an adjustment along the lines of Section 4 would ever be appropriate because the μ_i's cannot be exchangeable a priori. An appropriate analysis must take into account one's prior information concerning the relationships among the variables.

References

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Annals of Statistics, 2, 1152-1174.

Berger, J.O., and Berry, D.A. (1987). The relevance of stopping rules in statistical inference (with discussion). To appear in Statistical Decision Theory and Related Topics IV.

Berry, D.A. (1985). Interim analysis in clinical trials: Classical vs. Bayesian approaches. Statistics in Medicine, 4, 521-526.

Berry, D.A. (1987). Interim analysis in clinical trials: The role of the

likelihood principle. The American Statistician, 41, xxx-xxx.

Berry, D.A., and Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. Annals of Statistics, 7, 558-568.

Deely, J.J., and Lindley, D.V. (1981). Bayes empirical Bayes. Journal of the American Statistical Association, 76, 833-841.

Robbins, H. (1956). An empirical Bayes approach to statistics. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1, 157-163.