Statistical Science
2008, Vol. 23, No. 1, 45–47
DOI: 10.1214/08-STS236REJ
Main article DOI: 10.1214/08-STS236

Rejoinder: Microarrays, Empirical Bayes and the Two-Groups Model

Bradley Efron

The Fisher–Neyman–Pearson theory of hypothesis testing was a triumph of mathematical elegance and practical utility. It was never designed, though, to handle 10,000 tests at once, and one can see contemporary statisticians struggling to develop theories appropriate to our new scientific environment. This paper is part of that effort: starting from just the two-groups model (2.1), it aims to show Bayesian and frequentist ideas merging into a practical framework for large-scale simultaneous testing.

False discovery rates, Benjamini and Hochberg's influential contribution to modern statistical theory, is the main methodology featured in the paper, but I really was not trying to sell any specific technology as the final word. In fact, the discussants offer an attractive menu of alternatives. It is still early in the large-scale hypothesis testing story, and I expect, and hope for, major developments in both theory and practice.

The central issue, as Carl Morris makes clear, is the combination of information from a collection of more or less similar sources, for example from the expression levels of different genes in a microarray study. Crucial questions revolve around the comparability and relevance of the various sources, as well as the proper choice of a null distribution. Technical issues such as the exact control of Type I errors are important as well, but, in my opinion, have played too big a role in the microarray literature. The discussions today are an appealing mixture of technical facility and big-picture thinking. They are substantial essays in their own right, and I will be able to respond here to only a few of the issues raised.

I once wrote, about the jackknife, that *good* simple ideas are our most precious intellectual commodity. False discovery rates fall into that elite category. The two-groups model is used here to unearth the Bayesian roots of Benjamini and Hochberg's originally frequentist construction. In a Bayesian framework it is natural to focus on local false discovery rates, fdr(z), rather than the original tail area version Fdr(z). My apologies to Professor Benjamini for seeming to suggest that fdr is more immune than Fdr to correlations between the z-values. All false discovery rates are basically ratios of expectations, and as such remain relatively unbiased in the face of correlation. It is only the proof of the exact Fdr control property that involves some form of independence.

In the same spirit, I have to disagree that Fdr produces more reproducible results than fdr. Both methods operate at the mercy of an experiment's power, and low-power situations, such as the prostate cancer study, are certain to produce highly variable lists of "significant" cases. (At this point, let me repeat my plea for a better term than "significant" for the cases found to be nonnull, a dubious nomenclature even in classical settings, and definitely misleading for large-scale testing.)

As suggested by Figure 2, there is no great conceptual difference between fdr and Fdr, nor have I found much difference in applications. Table 1 says something about their comparative estimation accuracy. As Professor Cai suggests, the statistician can combine the two, using Fdr to select a reportable list of nonnull candidates, and fdr to differentiate the level of certainty within the list. Here the two roles reflect Benjamini's distinction between decision theory and inference, that is, between making a firm choice of nonnull cases and providing an estimate of just how nonnull they are.

As an enthusiastic collector of reasons to distrust the theoretical null distribution, I am happy to add *preselection of cases* to the list. Professor Benjamini correctly points out the dangers of this practice—among other things, it deprives the statistician of crucial evidence about the null distribution. If questioning the theoretical null seems heretical, it is worth remembering similar questions arising in classical ANOVA applications, for instance whether to use σ^2 (error) or σ^2 (interaction) in assessing the main effects of a two-way table. I share Benjamini's preference for finding the "right" theoretical null, but that is the counsel of perfection, often unattainable in examples like the education data.

Bradley Efron is Professor, Department of Statistics, Stanford University, Stanford, California 94305, USA (e-mail: brad@stat.stanford.edu).