EXPLORING THE INFORMATION IN P-VALUES

by

David Ruppert, Dan Nettleton, and J.T. Gene Hwang
Cornell University, Iowa State University, and Cornell University

November 9, 2005

ABSTRACT

A new methodology is proposed for estimating the proportion of true null hypotheses in a large collection of tests. The proportion of true null hypotheses is needed, for example, when controlling the false discovery rate in the analysis of microarray data. We assume that each test concerns a single parameter $\delta$ whose value is specified by the null hypothesis. Our methodology combines a parametric model for the conditional CDF of the $p$-value given $\delta$ with a nonparametric spline model for the density $g(\delta)$ of $\delta$ under the alternative hypothesis. The proportion of true null hypotheses and the coefficients in the spline model are estimated by penalized least-squares subject to constraints that guarantee that the spline is a density. The constrained, penalized least-squares estimator is computed efficiently using quadratic programming. Our procedure gives estimators with less bias compared to other estimators in the literature, which are positively biased because they are based upon an estimate of the marginal density of the $p$-values at 1. We define three estimators with different degrees of positive bias. In a simulation study, we compare our estimators to the convex, decreasing estimator of Langaas, Ferkingstad, and Lindqvist. The most biased of our estimators is very similar in performance to the convex, decreasing estimator. Our methodology produces an estimate $\hat{g}(\delta)$ of the density of $\delta$ when the null is false and can address such questions as “when the null is false, is the parameter usually close to the null or far away?” This leads us to define a “falsely interesting discovery rate” (FIDR), a generalization of the false discovery rate. We contrast the FIDR approach to Efron’s “empirical null hypothesis” technique. We discuss the use of $\hat{g}$ in sample size calculations based on the expected discovery rate (EDR). As an illustration, we analyze differences in gene expression between resistant and susceptible strains of barley after the plants have been exposed to a pathogen.