Simultaneously estimating many effect sizes, frequentist selection bias, and the bootstrap

Noah Simon*
University of Washington, Department of Biostatistics, Seattle WA – nrsimon@uw.edu

Abstract: With recent advances in high throughput technology, researchers often find themselves running a large number of hypothesis tests (thousands+) and estimating a large number of effect-sizes. Generally there is particular interest in those effects estimated to be most extreme. Unfortunately naive estimates of these effect-sizes (even after potentially accounting for multiplicity in a testing procedure) can be severely biased. In this talk we explore this bias from a frequentist perspective. We show that were the bias known apriori one could build estimates that (potentially significantly) dominate our usual estimators. In practice the bias will be unknown --- we discuss a bootstrap procedure to estimate it. Unlike other proposals for debiasing estimates, our procedure implicitly adjusts for unknown dependence between the features. Finally, we empirically demonstrate the efficacy of our approach and relate it to ideas in empirical Bayes and compound decision theory.

Keywords: high dimensional, selection bias, empirical bayes