Quality Preserving Databases: Statistically Sound and Efficient Use of Public Databases for an Infinite Sequence of Tests

Saharon Rosset

Professor
School of Mathematical Sciences
Tel Aviv University

Abstract:

Common data resources whose usage is open to the scientific community to facilitate research are becoming commonplace, especially in Biology and Genetics. The emerging scenario in which a community of researchers sequentially conduct multiple statistical tests on one shared database gives rise to major multiple hypothesis testing issues. It is often hard to control false discovery in the presence of unpredictable and sequential use, and existing tools are very limited.

We suggest a scheme we term Quality Preserving Database (QPD) for controlling false discovery without any power loss by adding new samples for each use of the database and charging the user with the expenses. The crux of the scheme is a carefully crafted pricing system that fairly prices different user requests based on their demands while controlling false discovery. The statistical problem encountered is one of defining appropriate measures of false discovery that can be controlled sequentially, and designing methodologies that can control them in the context of QPD. We describe a simple QPD implementation based on controlling the family-wise error rate using a method called alpha-spending, and a more involved implementation based on controlling a measure called mFDR, using an approach we term generalized alpha investing. We derive the favorable statistical properties of generalized alpha investing variants in general, and in the context of QPD in particular. The variant we implement can guarantee infinite use of a public database while preserving power, with very low costs, or even no costs under some realistic assumptions. We demonstrate this idea in simulations and describe its potential application to several real life setups.

A major concern in modern use of public datasets is adaptation (that is, when identity of future tests is based on the results of past tests). This is a concern which QPD does not address, and we survey more recent work on using variants of the well known "reusable holdout" framework for allowing both adaptiveness and indefinite usefulness through payment and data-increase schemes.

Joint work with Ehud Aharoni, Hani Neuvirth, Blake Woodward, Nathan Srebro and Vitaly Feldman.