

Multiple test problems and effective numbers of  
tests  
with applications to genetic association studies  
Prof. Thorsten Dickhaus  
(Humboldt Universität Berlin)

December 7, 2012

In many modern areas of science, several inferential problems have to be solved simultaneously on the basis of only one single dataset. Consider, for instance, genetic association analyses where many genetic loci are tested in parallel for association with a given phenotype. Mathematically, such problems can be formalized as multiple statistical hypotheses test problems and the data-analytic tools to solve them are multiple test procedures. The field of multiple hypotheses testing has become one of the major branches of mathematical and applied statistics during the past approximately 20 years, especially driven by the need for new data-analytic tools for problems from modern life sciences, such as genetics, proteomics, functional magnetic resonance imaging, or brain-computer interfacing.

After a general introduction into concepts for multiple hypotheses testing, we concentrate on a special class of multiple test problems, consisting of simultaneous point hypothesis tests in local statistical experiments. Under certain structural assumptions the global hypothesis contains exactly one element  $\vartheta^*$  (say), and  $\vartheta^*$  is least favorable parameter configuration with respect to the family-wise error rate (FWER) of multiple single-step tests, meaning that the FWER of such tests becomes largest under  $\vartheta^*$ .

Furthermore, it turns out that concepts of positive dependence are applicable to the involved test statistics in many practically relevant cases, in particular, for multivariate normal and chi-squared distributions. Altogether, this allows for a relaxation of the adjustment for multiplicity by making use of the intrinsic correlation structure in the data. We represent product-type bounds for the FWER in terms of a relaxed Šidák-type correction of the overall significance level and compute "effective numbers of tests".

Our methodology can be applied to a variety of simultaneous location parameter problems, as in analysis of variance models or in the context of simultaneous categorical data analysis. For example, simultaneous chi-square tests for association of categorical features are ubiquitous in genetic association studies. In this type of model, [Moskvina and Schmidt (2008)] gave a formula for an effective number of tests utilizing Pearson's haplotypic correlation coefficient as

a linkage disequilibrium measure. Their result follows as a corollary from our general theory and will be generalized.

The talk is based on [Dickhaus and Stange (2012)] and [Dickhaus et al. (2012)].

## References

- [Dickhaus and Stange (2012)] Dickhaus, T. and Stange, J. (2012). Multiple point hypothesis test problems and effective numbers of tests. SFB 649 Discussion Paper 2012-041. Sonderforschungsbereich 649, Humboldt-Universität zu Berlin, Germany.
- [Dickhaus et al. (2012)] Dickhaus, T., Straßburger, K., Schunk, D., Morcillo-Suarez, C., Illig, T. and Navarro, A. (2012). How to analyze many contingency tables simultaneously in genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, Volume 11, Issue 4, Article 12.
- [Moskvina and Schmidt (2008)] Moskvina, V. and Schmidt, K. M. (2008). On Multiple-Testing Correction in Genome-Wide Association Studies. *Genetic Epidemiology*, Volume 32, 567-573.