Final Program Report
Multiplicity and Reproducibility in Scientific Studies
Summer 2006

1. Objectives

Concerns over multiplicities in statistical analysis and reproducibility of scientific experiments are becoming increasingly prominent in almost every scientific discipline, as experimental and computational capabilities have vastly increased in recent years. This 2006 SAMSI summer program looked at the following key issues.

Reproducibility: A scientist plans and executes an experiment. A clinical trials physician runs a clinical trial assigning patients to treatments at random and blinding who has what treatment. A survey sampling person collects a survey. Scientists use statistical methods to help them judge if something has happened beyond chance. They expect that if others replicate their work, that a similar finding will happen. To clear a drug the FDA requires two studies, each significant at 0.05. There is growing concern that a number of such scientific studies are not replicating, and understanding the reasons for this was a top priority for the program.

Subgroup Analysis: Large, complex data sets are becoming more commonplace and people want to know which subgroups are responding differently to one another and why. The overall sample is often quite large, but subgroups may be very small and there are often many questions. Genetic data is being collected on clinical trials. Which patients will respond better to a drug and which will have more severe side effects? Disease, drug, or side effects can result from different mechanisms. Identification of subgroups of people where there is a common mechanism is useful for diagnosis and prescribing of treatment. Large educational surveys involve groups with different demographics, different educational resources and subject to different educational practices. What groups are different; how are differences related to resources and practices? What really works and why? Is the finding the result of chance? There is a need for effective statistical methods for finding subgroups that are responding differently. There is a need to be able to identify complex patterns of response and not be fooled by false positive results that come about from multiple testing. The program brought together statisticians and subject experts to develop and explore statistical strategies to address the subgroup problem. The benefit will be creditable statistical methods that are likely to produce results that will replicate in future studies.

Massive Multiple Testing: The routine use of massively multiple comparisons in inference for large scale genomic data has generated a controversy and discussion about appropriate ways to adjust for multiplicities. The program studied different approaches to formally describe and address the multiplicity problem, including the control of various error rates, decision theoretic approaches, hierarchical modeling, and probability models on the space of multiplicities, and model selection techniques. Applications include genomic data, clinical trial design and analysis, record matching problems, classification
in spatial inference, and anomaly discovery and syndromic surveillance. The goal of the program was to identify the relative merits and limitations of the competing approaches, and to understand which features of reproducibility are addressed.

2. Workshops

2.1 Opening Workshop

The Kickoff Workshop was held July 10-12, 2006, and focused on clear formulation of the challenges in the area, especially the issues of reproducibility of scientific experiments, massive multiple testing, and subgroup analysis. The workshop set the stage for the subsequent Program research, through tutorials, talks and discussion sessions designed to focus attention on the key problems to be considered by the program working groups.

2.2 Transition Workshop

An informal Transition Workshop was held July 27-28, 2006, summarizing the results of the Program research and discussing the remaining challenges.

3. Working Groups and Outcomes

3.1 Report on Subgroup Analysis working group

The activities of this working group combined consideration of background papers, general methodological talks, and related discussions with emphasis on practical advice to researchers on how to handle multiplicity in their studies and particularly how to report results. The focus of this advice was primarily on epidemiological studies and clinical trial research in medicine.

Much of our background preparation involved becoming acquainted with two types of published articles:

(a) the many publications giving advice on how to (or whether to) carry out subset analyses, whether planned or unplanned, in such research, taking into account the problems of possible inflation of errors due to multiple testing. Many of these articles give good advice, but many such studies are still being conducted without following that advice.

(b) articles documenting the large number of reported results of clinical trials and epidemiological studies that are not confirmed in follow-up research, many of which include subgroup analyses. It seems clear that a major contributor to this lack of confirmation is inattention to, or inadequate appreciation of, the effects of multiple testing.
Our group consisted of both Bayesians and frequentists. While we may have had differences in the ways in which we approach problems, we had broad areas of agreement on the basic issues.

We had some preliminary discussions about types of subgroups and subgroup analyses. One way of dividing such analyses is in terms of whether they concentrate on (i) a set of demographically-defined subgroups, e.g. groups defined in terms of race-ethnicity, geography, genomic classification, etc. to see which show either effects of a treatment or effects that differ from those of other such subgroups, or (ii) what might be called outcome-based analysis: classifying patients into groups based on their treatment outcomes, and looking for demographically defined subgroups or covariates that predict the differences, e.g. in a medical study looking at subgroups of patients in which the treatment is effective, not effective, harmful. Studies can focus on one of these aspects or combine them.

The former approach often starts with a small number of defined groups, although genomic analysis has recently led to much larger numbers. In this context subgroup analysis sometimes involves outcomes separately within each subgroup, whether a treatment does or does not have apparently-significant effects in the total group, and sometimes comparing sizes of effects in different subgroups. The latter usually considers large numbers of covariates. The former type often uses methods such as analysis of variance and regression; the latter often involves clustering methods. Some of the talks in our session fit into each of these approaches; e.g. Robert Obenchain's talk dealt mostly with the outcome-based analysis while Siva Sivaganesan took primarily the defined subgroup approach.

We appreciate the fact that in complex studies it is not easy to plan analysis and reporting taking multiplicity into account. Rather than proposing specific methods, or specific types of error control (e.g. control of family-wise error, false discovery rates, consideration of posterior probabilities, etc.) which in any case would depend on many characteristics of the studies, we decided to propose a three-level plan for taking multiplicity into account that would work with a variety of specific types of studies, types of error control, and details of analytic plans. In fact, our advice is general enough to cover all types of tests, not only those based on subgroup analysis.

Recognizing that many researchers will explore their data and are anxious to glean as much as possible from the results, we propose some compromises between strict adherence to pre-determined protocols and unlimited data dredging in reporting study results. We suggest that researchers dedicate a 5% significance level for testing a small, targeted group of primary outcomes (perhaps only one) and an additional level, possibly 5%, for testing an entire set of pre-designated secondary outcomes. Any additional outcomes that draw attention, either because of very small p-values or large estimated effect sizes, should be reported as speculation, unsupported by good statistical methods even if those p-values or estimated effect sizes appear to be unusually compelling. In some cases it may be possible to estimate an overall error probability for this exploratory set, but in most cases it will not be possible. Any theory or empirical evidence supporting these additional results should be presented, but the results remain speculation unless supported by additional targeted research.

We propose that the level at which secondary outcomes are tested should be at most the original level. One advantage of using the original level is that if some
secondary outcomes are very highly correlated with the original ones, there is no loss of power in using that level, assuming appropriate consideration is given to the correlations within the secondary class.

We made plans for a joint paper or papers with our recommendations and with detailed information, general analytic advice, and supporting discussion of concrete examples. No further progress has been made on this paper to date, but a number of activities, stimulated by our discussions and presentations, and aided greatly by the note-taking of Rochelle Trachtenberg, have been undertaken by participants in the workshop.

Some activities that have been influenced by the program:

1. Siva Sivaganesan, Prakash Laud and Peter Mueller continued work on a joint project on a Bayesian approach for subset analysis in clinical trials. Discussions in the workshop provided important background and motivation for continued work on this project. The proposed approach is on the interface of the two workshop themes "subset analysis" and "multiplicities." Besides the discussions in the subgroup working group we also gained important insights for this project from interactions with colleagues in the multiplicity working group. Results have been presented at the recent ENAR meeting. A draft manuscript is in preparation.

2. Stanley Young and Juliet Shaffer co-sponsored a three-hour symposium at the February 2007 meeting of the American Association for the Advancement of Science entitled "Mixed health messages: Observational versus randomized trials" with five speakers and three discussants. Juliet Shaffer spoke on "History of multiple testing" with much content informed by workshop discussions. (The workshop was well-attended with lively ensuing discussion.)

3. Juliet Shaffer was a participant in the Grants Award Conference sponsored by the American Educational Research Foundation from Sept. 28-Oct. 1, 2006. This conference was to give research advice to graduate students and new PhDs who had received AERA research grants. Juliet had been involved in preparation of the ASA-NSF-sponsored report "Using Statistics Effectively in Mathematics Education Research" and gave a report on this guide, also emphasizing issues covered in the SAMSI workshop, especially noting the importance of replication and results of replications of medical studies. (In discussion following the report, students noted that they are often discouraged from replicating studies, informed that "prestige" requires innovative research.)

4. At the Third Erich L. Lehmann symposium, to be held in May 2007, there will be two sessions related to and influenced by the workshop: (a) a session organized by Jim Berger on the workshop in general, and (b) a session organized by Juliet Shaffer on multiplicity issues related to the workshop topics. One of the speakers in the latter session will be Charles Lewis, who will speak about similar multiplicity issues in education and psychology, areas for which the workshop material is relevant but which were not addressed there.
5. The workshop influenced Juliet Shaffer to work on some multiple comparison issues in clustering which is now ongoing with two psychologists, Harvey Keselman and Rhonda Kowalchuk. Some results will be reported at the Fifth International Conference on Multiple Comparison Procedures to be held in July, 2007.

6. Bob Obenchain (Eli Lilly) has continued his "meaningful subgroup" research on adjustment for all forms of bias in observational (nonrandomized) studies. Through systematic application of sensitivity analyses, Bob's "Local Control" (LC) approach forms, splits, compares and (following over-shooting) re-combines subgroups of most comparable patients. This approach has considerable appeal to Bayesians because, without imposing any prior distribution, it reveals and smooths the sample distribution of local treatment differences (LTDs.) Rather than focus on point estimates in very large "samples" and their highly questionable p-values, Bob's approach uses a simple nested ANOVA model (treatment within patient cluster) and forms non-parametric confidence and tolerance intervals or regions. Bob presented an update of his LC concepts at the March 2007 ENAR meeting and is also working with Doug Faries (Lilly), Andrew Leon (Cornell) and Josep Maria Haro (Barcelona) to publish a book on "Analysis of Observational Health-Care Data."

3.2 Report on Massive Multiple Testing (MMT) working group

The plan for this working group was to study different approaches, consider applications in inference for genomic data and other research problems that require massive multiple testing. The goal was to identify the relative merits and limitations of competing approaches for diverse applications.

3.2.1 Workshop, Lectures and Group Discussions

Speakers and topics for the talks in the opening workshop were chosen to reflect the diversity of approaches proposed in the recent literature. Different approaches that were discussed included model-based Bayesian approaches, frequentist false discovery rate control, decision theoretic approaches and theoretical considerations of the false discovery process as a function of a threshold for the rejection region. Literature related to these approaches has developed largely separately with most researchers working exclusively in one research direction. The workshop provided an opportunity for all participants to learn about current problems and ideas across different approaches.

The workshop program intentionally mixed talks in the three big program areas, reproducibility, subgroup analysis, and massive multiple testing. Although all three areas depend on common underlying mathematical principles, recent research in these areas has developed separately. Similar to the exchange among researchers working in different research areas within multiple testing, exchange of ideas across the three working groups lead to interesting new applications and research projects reported below.

3.2.2 Working Group Discussions and Results
After the opening workshop the MMT working group continued to meet almost daily, usually inviting one participant to present an in depth discussion of a selected research topic. This mechanism and informal exchange among participants working in the different working groups naturally lead to interesting new research projects. We report some examples that we are aware of.

**Loss function for credible intervals:** In informal discussions between Dani Yekutieli and Ken Rice the question of a formal decision theoretic justification of Bayesian credible intervals was brought up. This was in reaction to earlier talks that pointed out various derivations of FDR-based rules as Bayes rules on one hand, and a review of methods to control FDR-type error rates for frequentist confidence intervals on the other hand. Despite the routine use of credible intervals in Bayesian data analysis, it seemed there was no good reference to derive them as formal Bayes rules. In following research Ken Rice developed an appropriate argument. This has meanwhile led to a draft manuscript that will likely become a classic reference.

**Bayesian justification of FDR rules:** Another discussion involved Yoav Benjamini, Ken Rice and others. In one of the daily lectures S. Bayarri conjectured that it was not possible to justify FDR control as a Bayes rule from first principles, without including realized FDR as an explicit part of the loss function. Several alternatives were discussed that approximately lead to rules similar to Benjamini and Hochberg’s FDR control. Ken Rice has meanwhile developed an argument based on a loss function for credible intervals. The rule approximately leads to FDR control.

**Bayesian subgroup analysis and edge detection:** A graduate student participant, James Scott, reports that the workshop motivated three projects that will likely constitute part of his thesis. The projects are fully Bayesian treatment of subgroup analysis, variable selection with multiplicity control, and multiplicity selection for edge detection in spatial inference.

**Gene-environment interactions:** Woncheol Jang has developed two multiple testing projects as a result of his workshop participation. In one project he will consider models with gene-environmental interactions that include multiplicity control. In the second project he will consider multiple testing for high dimensional data as they arise from microarray experiments.

**Bayesian subgroup analysis:** Siva Sivaganesan, Prakash Laud and Peter M’uller developed an approach for subgroup analysis based on Bayesian multiplicity control. The project is also reported as an outcome of the subgroup analysis working group.

H. Finner and T. Dickhaus report the following research nuggets:
- **FDR control**: Discussions with Yoav Benjamini, Daniel Yekutieli and Sanat K. Sarkar on possible improvements of FDR-controlling procedures.

- **Dependence**: Discussions on volatilitity of the FDR proportion under dependency. Gene expression data: Learning about genome-wide association studies / SNP data analysis (discussion of Lei Sun)

- **Asymptotics**: Large p, small n asymptotics

- **New projects**: Two new projects were initiated, “Asymptotic improvements of some FDR procedures based on an asympototically optimal rejection curve” and “Optimal rejection curve for FDR control in various models.”

### 3.2.3 Conference and Workshop Presentations

Participants at the SAMSI workshop have organized sessions in several prominent statistics meetings that will highlight results from the SAMSI workshop. Stan Young set up a 3 hour invited session at the annual meeting of the AAAS (American Association for the Advancement of Science). An attending reporter for the Economists wrote a short article that appeared in the print edition of the Economist.

Jim Berger and Peter Müller organized a session at the ENAR (Eastern North America Region of the Biometric Society) meeting in Atlanta, with talks related to subgroup analysis and multiplicity control. The session was well attended and highlighted results from the SAMSI workshop.

Jim Berger and Juliet Shaffer are organizing two sessions at the upcoming Lehmann Symposium at Rice University, Houston. The diverse nature of the AAAS, ENAR and Lehmann Symposium meetings reflects the universal importance of the SAMSI program, and the diversity of the program participants.

### 3.2.4 Manuscripts

The following papers will acknowledge SAMSI support.

- Ken Rice, ”Bayesian Decision Theory for Multiple Comparisons”

- Siva Sivaganesan, Prakash Laud Peter Mueller, ”Subgroup analysis - a Bayesian decision theoretic approach”

  Ken mentions:
  
  ”Draft title above, I am working on the draft paper right now! It will say extremely nice things about SAMSI in the acknowledgements, I wouldn’t be working on this without the workshop opportunity.”

H. Finner and T. Dickhaus report that the workshop motivated them to write up the following paper:

- H. Finner, T. Dickhaus And M. Roters, “on the false discovery rate and an asymptotically optimal rejection curve,” submitted for publication.