Organizing Committee

Chair: **Dr Ernest Fokoué**, Rochester Institute of Technology

Dr Tanzy Love, University of Rochester
Dr Bruce Blaine, St John Fisher College
Ms Linda Battaglia, Harris Interactive
Ms Georgette Nicolaides, Syracuse University
Ms Jennifer Bergamo, SUNY-OCC
Dr Jen-Ting Wang, SUNY-Oneonta

Keynote Speaker

Professor Jason Hsu
Department of Statistics,
The Ohio State University
Columbus, OH 43210
eMail: jch@stat.osu.edu

Conference Venue

Rochester Institute of Technology, Rochester, NY
Building 9 – College of Engineering Building
(a) **Xerox Auditorium** (b) **Room 2159** (c) **Room 2149**

Conference Sponsors

American Statistical Association (ASA) – **Rochester and Syracuse Chapters**
Center for Quality and Applied Statistics – **Rochester Institute of Technology (RIT)**
Rochester Institute of Technology (RIT)

Conference Program at a glance

9:30-10:00am	**Welcome, Orientation and breakfast**
10:10-10:55am	Post-breakfast sessions (2 parallel sessions)
11:10-11:55am	**Pre-lunch sessions (2 parallel sessions)**
12:00-1:00pm	Lunch (sandwiches served out Xerox Auditorium)
1:15-2:15pm	**Keynote lecture (Xerox Auditorium)**
2:30-3:15pm	**Post lunch sessions (2 parallel sessions)**
3:30-4:15pm	Final oral session (Xerox Auditorium)
4:30-5:00pm	**Poster session (Presenters set your posters in Room 2149 as soon as possible)**
5:00-5:30pm	Wrap-up session (Xerox Auditorium)

Post Breakfast - Session A — Xerox Auditorium, Building 9

Interesting Bayesian and Frequentist Applications of Statistics

Session chair: Ms Jennifer Bergamo, SUNY-OCC, NY

10:10-10:30am Effect Modification using Latent Mixture Analysis

Dr Tanzy M. T. Love, Dr Sally W. Thurston and Dr Philip W. Davidson, University of Rochester
eMail: tanzy_love@urmc.rochester.edu

The Seychelles Child Development Study (SCDS) is examining associations between prenatal exposure to low doses of methylmercury (MeHg) from maternal fish consumption and children's developmental outcomes. Whether MeHg has neurotoxic effects at low doses remains unclear and recommendations for pregnant women and children to reduce fish intake may prevent a substantial number of people from receiving sufficient nutrients that are abundant in fish. The primary findings of the SCDS are inconsistent with adverse associations between MeHg from fish consumption and neurodevelopmental outcomes. However, whether there are subpopulations of children who are particularly sensitive to this diet is an open question. Secondary analysis from this study found significant interactions between prenatal MeHg levels and both caregiver IQ and income on 19 month IQ [Davidson et al., 1999]. These results are dependent on the categories chosen for these covariates and are difficult to interpret collectively. In this paper, we estimate effect modification of the association between prenatal MeHg exposure and 19 month IQ using a general formulation of mixture regression. Our model creates a latent categorical group membership variable which interacts with MeHg in predicting the outcome. We also fit the same outcome model when in addition the latent variable is assumed to be a parametric function of three distinct socioeconomic measures. Bayesian MCMC methods allow group membership and the regression coefficients to be estimated simultaneously and our approach yields a principled choice of the number of distinct subpopulations. The results show three different response patterns between prenatal MeHg exposure and 19 month IQ in this population.

10:35-10:55am Bitemarks and the Product Rule: A cautionary tale of statistics in the forensics sciences

Dr H. David Sheets, Department of Physics, Canisius College, Buffalo, NY
eMail: sheets@canisius.edu

Strong claims have been made about the effectiveness of forensic bitemark analysis as used in criminal cases. One study in particular has presented a statistical analysis claiming to support low enough probabilities of random matches in the human anterior dentition to make the biting dentition effectively unique to individuals. Recent criticisms of bitemark analysis by the National Academy of Sciences and a number of false convictions based on bitemarks has prompted a closer look at these statistical claims. The use of a uniform distribution model and the product rule (implying uncorrelated data) in the published model prompts us to question the validity of the claimed uniqueness of the human biting dentition. This case has possible pedagogical use, as an illustration of the importance of checking the assumptions of a statistical model against the properties of the underlying data.

American Statistical Association

Syracuse Chapter

Post Breakfast – Session B — Room 2159, Building 9

Recent advances in Statistics and Data Mining

Session chair: Dr Jen-Ting Wang, SUNY-Oneonta, NY

10:10-10:30am

Combining various Kernels for Efficient Pattern Extraction from Medical Records

Mr Edwin Anto, Department of Computer Science, Rochester Institute of Technology, NY
eMail: exa6414@rit.edu

In this talk, we present a brief overview on Text Mining and a preliminary exploration of pattern extraction from medical records from a geriatrics practice at Rochester General Hospital. We propose to combine elements of modern day text mining with traditional data mining to build both unsupervised and supervised learning models to help understand various diseases in patients beyond the age of 61. The results of this study can help identify better and early detection of diagnosis for future patients based on the similarity of their conditions to past cases.

10:35-10:55am

A Geometric Illustration of Differential Gene Association Analysis for Microarray Data

Dr Rui Hu, Department of Biostatistics and Computational Biology, University of Rochester, eMail: Rui_Hu@urmc.rochester.edu

Microarray gene expression analysis has become a routine in medical research in the past decade. A common and important question in gene expression data analysis is the identification of biologically "interesting" genes. Recently, we have developed a differential gene association analysis to select genes based on the change of their dependence structures across different phenotypes or biological conditions. A geometric illustration is provided in this paper to help understand the changes of inter-gene dependence structures and how this information can be used to detect biologically meaningful genes.

American Statistical Association

Rochester Chapter

http://amstat.org

Pre Lunch – Session A — Room 2159, Building 9

Variable selection and model selection

Session chair: Dr Tanzy Love, University of Rochester, Rochester, NY

11:10-10:30am

Model Ranking Applications in Paleoecology

Dr John C. Handley, Rochester Academy of Science, Rochester, NY USA
eMail: jhandley@rochester.rr.com

Information theoretic model ranking is an often-used paradigm for statistical analysis in ecology and paleoecology. The basic concept is that several models are posited to explain data and the best one is chosen by how well the model approximates an unknown true model. Model ranking in an alternative to hypothesis testing. It provides a simpler way to analyze data from scientific studies and is less prone to misinterpretation than
hypothesis testing.

11:35-11:55am

The Controversy of Variable Importance in Random Forests

Dr Padraic Neville, Fairport, NY and SAS, Cary, NC
eMail: Padraic.Neville@sas.com

A good measure of the importance of a variable in a model helps prioritize which variables to examine further. The variable importance from a random forest model of genomics data can suggest which genes are likely to interact with other genes because variables contribute to the random forest model through interactions. Two measures of importance have been proposed for random forests: impurity-reduction and error-increase. Impurity-reduction is used in CART decision trees and in gradient boosting machines. Error-increase is usable but not used in practice. Leo Brieman (2001) proposed error-increase for forests, but later deemed it unreliable (2003). Richard Berk (2003, 2008) says that only error-increase should be used. Among several recent papers in Bioinformatics studying the reliability of error-increase, Nicodemus and Malley (2009) present unchallanged evidence preferring error-increase to impurity-reduction. Van de Laan rejects both measures (2006). This presentation defines random forests and the two measures of variable importance, and illustrates how using either measure random forests can find better candidates for gene interactions than selecting variables with t-test or LASSO. The arguments of Berk and Nicodemus and Malley against impurity-reduction are shown to be flawed because they compute impurity-reduction with training data and error-increase with hold-out data. The relative value and reliability of the two measures of variable importance in random forests remains unresolved.

Pre-lunch – Session B — Xerox Auditorium, Building 9

Approaches to Teaching Core concepts

Session chair: Dr Bruce Blaine, St John Fisher College, Rochester, NY

11:10-11:30am A Semester-Long Project to Engage Students in Elementary Statistics

Dr. Jennifer Bready, Mount Saint Mary College, Newburgh, NY
eMail: Jen.bready@msmc.edu

In this article the author will discuss a project used in her elementary statistics class. This project begins on the first day of class and culminates at the end of the semester in a statistical research paper. Students are involved in every step of the process, from choosing a topic and title, writing questions, collecting data, and analyzing results.

11:35-11:55am Presenting Statistical Concepts To Non-Technical Audiences

Ms Bahati Benjamin, Symmetry Analytics, LLC
eMail: info@symmetrystats.com

Statistics is often daunting and somewhat of a mystery to the non-technical learner. This paper provides a framework from which to translate mathematical concepts to everyday experiences. At Symmetry Analytics, we believe that mathematics is an essential part of life. The Stats Museum was created to explore the beauty and wonder of the mathematical sciences. The museum utilizes an online platform to explore the elements of Success using box plots, hierarchical graphs and Venn diagrams. Similarly, the Women exhibit explores every day experiences of women including love, time management and triumphs utilizing type II error comparisons, line graphs and other graphical techniques. The exhibits are constructed to be welcoming and engaging for non-technical audiences. However, much of the content is relevant and pertinent to statisticians who communicate with diverse participants.

Post-lunch – Keynote Lecture — Xerox Auditorium, Building 9

1:15-2:15pm **Professor Jason Hsu** The Ohio State University, Columbus, OH eMail: jch@stat.osu.edu

Title: So, What Is Personalized Medicine?

What examples of Personalized Medicine do you know? One form of personalized medicine is: Order kit, send in saliva sample, open wallet, select which defects to be tested for, wait for results. How reliable are such tests? Are they FDA-approved? Another form is a drug targeting a subgroup of the patient population. After describing examples in oncology and cardiology, we will try to resolve some not easily recognized statistical issues in Genome-wide Association Studies (GWAS), studies that aim to discover association between biomarkers and diseases. For example, in biomarker discovery studies, controlling the false discovery rate (FDR) is popular. Suppose FDR is set at 5%, and 1000 out of one million markers are found to be significantly associated with a disease. Is the interpretation "50 out of the 1000 discoveries are expected to be false" appropriate? As another example, instead of cross-validation or external validation, permutation tests are often used to validate biomarker discoveries. Is it true that permutation multiple testing builds correct null distribution without model assumption? You might think that I have posed these issues in the form of quiz questions because the theme of this conference is statistical education ☺ Let us see how simple or tricky these questions are!

Post lunch – Session A — Xerox Auditorium, Building 9

Bayesian Solutions to Frequentist Challenges in Basic Statistics

Session chair: Ms Georgette Nicolaides, Syracuse University, Syracuse, NY

2:30-2:50pm

How do college science faculty (mis)interpret nonsignificant p values?

Dr Bruce Blaine, Statistics Department, St. John Fisher College, Rochester, NY
eMail: bblaine@sjfc.edu

Amid the broader criticism of the use of null hypothesis signficiance testing (NHST), much research has shown that students and faculty misunderstand significant *p* values from significance tests. Less is known about how people interpret nonsignificant *p* values. Some research finds that people conflate nonsignificant *p*s with evidence for H_0, whereas other research shows that nonsignificant *p*s are "spun" to reflect support for H_1. College faculty in the natural and social sciences were surveyed on their interpretations of a nonsignificant *p* value from an experiment. Survey items addressed the perceived implications of $p>.05$ for H_0, H_1, and replication probability. The data revealed confusion about what nonsignificant *p* values meant, and some evidence that misinterpretations of nonsignificant *p* values were related to the underestimation of sampling error.

2:55-3:15pm

To Bayes or not to Bayes? How the Bayesian paradigm can help boost students' understanding of Statistics

Dr Ernest Fokoué, Center for Quality and Applied Statistics, Rochester Institute of Technology
eMail: ernest.fokoue@rit.edu

Despite their widespread use, frequentist concepts such as P-values and confidence intervals are not as intuitive as many users might unsuspectingly believe. In fact, one would be quite justified in thinking of them as backward. It is very common to have students mistake confidence for probability, not realizing that frequentist confidence intervals make statements about hypothetical samples and not about the sample at hand. The Bayesian counterparts of frequentist concepts, although deemed more demanding mathematically due to the extra layers of knowledge they require, turn out to be more intuitive, and even provide a more natural support for understanding fundamental statistical ideas. In this talk, I will point to various ways in which the Bayesian paradigm has the potential to make it easier for students to fully grasp most statistical inference concepts.

Center for Quality and Applied Applied (CQAS)

Visit us today

http://www.rit.edu/~w-sgas/

Post lunch - Session B — Room 2159, Building 9

Teaching Experiments and Medical Applications

Session chair: Dr Daniel Lawrence, Rochester Institute of Technology, Rochester, NY

2:30-2:50pm Statistics in medical device industry - a mini review

Ms Jiejing Qiu, Clinical, Medical & Scientific Affairs, Welch Allyn, Skaneateles Falls, NY
eMail: jiejing.qiu@welchallyn.com

Statistics is broadly applied in both development and validation processes in medical device industry. It plays critical roles in study design, data analysis and results interpretation to ensure the safety and effectiveness of medical devices and compliance with FDA regulations.. In this presentation, a Biostatistician's roles and responsibilities in a medical device company will be reviewed. Specifically, we will discuss some common statistical tools used in these studies, as well as the challenges and the opportunities that exist in medical device related clinical studies. This topic may be useful for promoting collaborations among statisticians in industry and academia. It may also be useful for students who are interested in working in medical device industry.

2:55-3:15pm Experiment with Experiments

Dr Joseph Voelkel, Center for Quality and Applied Statistics, Rochester Institute of Technology
eMail: joseph.voelkel@rit.edu

The value of experiential learning is illustrated with a variety experiments that have been designed, executed and analyzed by students in an applied course in experimental design. The gap between classroom examples and actual experiments is illustrated with a number of entertaining examples, with the hope of encouraging statistics instructors to "experiment with experiments."

Rochester Institute of Technology (RIT)

Final session - Xerox Auditorium, Building 9

Special Applications and Aspects of Survey Analysis

Session chair: Ms Linda Battaglia, Harris Interactive, Rochester, NY

3:30-3:55pm Digging in Unburried Treasure: Analysis of Samples from the "Glass Wreck" of Serce Limani (ca. 1020 A.D.)

Mr Dean Neubauer, Corning Incorporated, Corning, NY

eMail: neubauerdv@corning.com

It is believed that a ship carrying glass artifacts and cullet (recycled glass) sank off the coast of Serce Limani, Turkey ca. 1020 AD. Investigating the nature of the Serce Limani glass will provide important information for future research on medieval glass, glass production, and trade. Many articles have been published over the last 25 years on the nature of this glass, but compositional analysis has been preliminary. The Corning Museum of Glass (CMoG) has analyzed some 400 Islamic glass samples. Most of these samples are from fragments taken from 20 different archeological sites (n=180 samples have been used for this analysis). The data present a general picture of chemical compositions of medieval Islamic glass which provides a benchmark for identification of the origin of the Serce Limani glass, and may provide clues as to where the ship originated prior to its demise. This is the fundamental problem that this paper attempts to answer is—where did this ship originate based on the compositional analysis of its glass contents? For over 25 years scientists have tried to extract this information but without statistics! The problem is framed initially from the glass researcher viewpoint, and then multivariate methods are used in succession to gradually extract the information contained therein to identify the probable origin of the glass artifacts.

4:00-4:25pm Are survey responses unidimensional metrics? A case for multidimensionality

Dr Daniel R. Lawrence, Rochester Institute of Technology, NY

eMail: drleas@rit.edu

Even though survey data are usually categorical in nature, the chosen method of analysis often assumes the data to be ratio (or at least interval). The rationale for this choice is sometimes simply that the results of such analyses are more easily explained and/or the software packages for doing those analyses more readily accessible. Just as troubling is the implicit assumption that survey respondents think along the same continuum, implying that the data are unidimensional, thus calling for an analysis such as a single-dimension multiple-regression analysis (MRA). In fact, people don't think alike, and as a consequence, any collection of the measures of opinions or preferences will very likely have some sort of latent multidimensional structure. For example, in a satisfaction survey, a set of Overall Satisfaction ratings will probably include a few low or even very low ratings, some moderate ratings, and (hopefully) a few high ratings. It is both unreasonable and unrealistic to assume that the same action items (or "drivers") for effecting a shift in responses from Very Dissatisfied or Dissatisfied, say, up to Satisfied would be the same items as those that should be the focus for moving respondents who are presently Satisfied up to a level of Very Satisfied. Action items associated with those two distinct shifts in attitude will almost certainly be different. That in mind, a proper method of analysis should be used, one that recognizes the data for what they are—namely, categorical with a multidimensional structure.