Organizing Committee

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

Dr Tanzy Love, University of Rochester
Ms Katie Evans, University of Rochester

Keynote Speaker

Professor George Michailidis
Department of Statistics,
The University of Michigan
Ann Arbor, MI
eMail: gmichail@umich.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)**
Journal of Unified Statistical Techniques (JUST)
PRAXAIR

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)**
3:30-4:25pm	Final oral session (Xerox Auditorium)
4:30-5:00pm	Wrap-up session (Xerox Auditorium)

Conference Website

http://asaupstateconference.syr.edu

Post Breakfast - Session A — Xerox Auditorium, Building 9

Tutorial - Elements of Regression Analysis

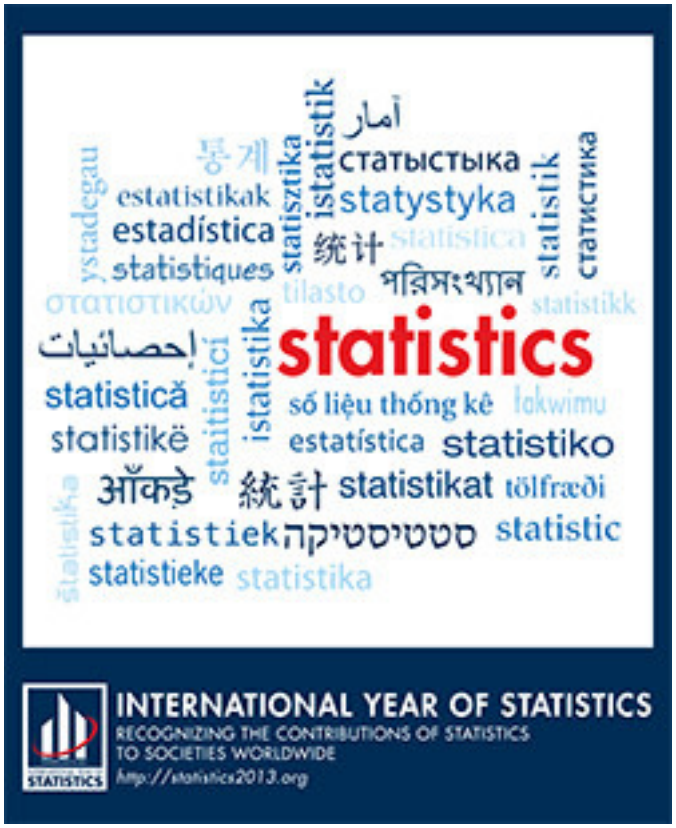
Session chair: Prof Joseph Voelkel

10:10-10:55am **Ridge Regression Estimators**

Professor Marvin Gruber, Rochester Institute of Technology, Rochester, NY
eMail: mjgsma@rit.edu

We will tell what ridge regression estimators are and illustrate how they may be helpful for the analysis of multicollinear data. How to formulate the estimators and use them will be illustrated by a specific example.

The origin and derivation of the ridge estimators will be taken up from three points of view. These will include a frequentist approach, a Bayesian approach and as a special case of a problem about linear operators in a Hilbert space. We will conclude by telling how two groups of researchers might learn from one
another and benefit greatly by working together.



http://www.statistics2013.org/

Post Breakfast - Session B — Room 2159, Building 9

Multifaceted Applications of Statistical Science

Session chair: Dr Yusuf Bilgic

10:10-10:30am **A Statistical Determination of the Characteristics of Playoff Teams in the NHL**

Dan Foehrenbach, Rochester, NY
eMail: dfoehrenbach@gmail.com

This paper uses machine learning and multivariate techniques to explore different team performance measurements used in the National Hockey League. Every aspect of hockey (offense,defense,special teams and goal-tending) is used within this analysis. The main goal is to determine which characteristics are most responsible for the success of teams. Most of the statistics and success indicators are commonly used like goals per game, power play percentage, playoff appearance and championship wins. In this paper, I seek to establish which aspects of hockey generate the most playoff teams. Using 11 years worth of data from 2000 to 2012 the application of techniques such as exploratory data analysis,cluster analysis, logistic analysis, support vector machine and linear discriminant analysis reveal compellingly interesting and consistent (over the years) elements of success. Lastly, a prediction will me made for the teams to make the playoffs for the current shortened NHL season.

10:35-10:55am **Designing a genome-based HIV incidence assay with high sensitivity and specificity**

Dr Tanzy M. T. Love, University of Rochester, Rochester, NY
eMail: tanzy_love@urmc.rochester.edu

Considerable inaccuracy in identifying HIV incidence has been a serious obstacle to the development of efficient HIV/AIDS prevention and interventions. Accurately distinguishing recent or incident infections from chronic infections enables one to monitor epidemics and evaluate the impact of HIV prevention/intervention trials. Our study designed a novel scheme for identifying incident infections in a highly accurate manner, based on the characteristics of HIV gene diversification within an infected individual. We devised a binary classification test based on the tail characteristics of the Hamming distance distribution of sequences and identified a clear signature of incident infections. The presence of closely related strains in the sampled HIV gene sequences identify recent infections in both single-variant and multivariant transmissions. This criteria, used as a biomarker, is found to have greater than 95% specificity and sensitivity and is robust to viral and host-specific factors. Because of rapid and continuing improvements in sequencing technology and cost, sequence-based incidence assays hold great promise as a means of quantifying HIV incidence from a single blood test.

Pre Lunch - Session A — Xerox Auditorium, Building 9

Tutorial – Introduction to Sports Statistics

Session chair: Mr Padraic Neville

11:10-11:55am

Introduction to Sports Statistics

Dr. Chulmin Kim, Rochester Institute of Technology, Rochester, NY 14623
eMail: cxksma@rit.edu

Introduction to Sports Statistics to use examples in major league baseball, college basketball and college football Chulmin Kim Rochester Institute of Technology

Most sports events adopt numerical scoring system. One point in soccer means a goal, two or three points in basketball means a field goal, and six points in football means a touchdown. Some sports events like marathon and long distance speed skating use the rank and the time measurement. Players are sometimes evaluated their performance by statistics. In professional sports, their statistics may be a big factor to determine their salary. People fascinate the sports because of their uncertainty of winning. Sports coaches and leaders may analyze their players' statistics in addition to the research from psychology and medical sciences not only to improve the individual performances but also to obtain better winning chance. Nowadays more people like to manage their own teams heavily based on the real sports statistics in fantasy sports games and we can learn the importance of statistical mind through the film like "Money Ball (2012)". We illustrate how statistical thinking can be used to answer some interesting questions in major league baseball, college basketball and college football

Pre Lunch – Session B — Room 2159, Building 9

Educational Pointers

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

11:10-11:30am **Trends and updates in the teaching of inferential statistics**

Dr Yusuf Bilgic, SUNY Geneseo, NY
eMail: bilgic@geneseo.edu

Since the 1990's, there has been an ongoing discussion on how statistics education should be reformed in order to enhance its effectiveness. Recently, many reform movements and resources in statistics education have begun to reshape both the content of the K-16 statistics curriculum and the methods that we use to teach that material. Examples include the GAISE Recommendations, the CATALYST Project, the Cause Organization, and Project MOSAIC. One of the major directions of the statistics education reform movement involves changes in the philosophical content. For example, the Fisher, Neyman-Pearson, maximum likelihood, and Bayesian approaches to hypothesis testing all emphasize different aspects of the statistical testing/inference process.

In this talk, I will discuss this historical trend in the teaching of inferential statistics and consider the position of hypothesis testing and the use of p-value in current and developing curricula

11:35-11:55am **Another Look at Design of Experiments**

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

Through several examples, we examine some key principles of designed experiments, and what can happen if these principles are not followed. These principles include methods to reduce bias; to increase precision; and to provide an objective method for detecting differences—for example, whether a treatment (a new drug, say) is better than a control (the best available drug on the market, say). Our presentation will include ideas such as randomization, replication, and blocking; single- and double-blinding; factorial structures; and covariates

.

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

1:15-2:15pm

Professor George Michailidis The University of Michigan, Ann Arbor, MI
eMail: gmichail@umich.edu

Title: **Statistical Methods and Issues in the Analysis of Network Data**

Over the last few years, researchers in diverse fields have collected a large number of network data. As a result, statistical models are introduced and methods developed for their analysis and inference.

In this talk, we examine some canonical problems including network inference from high dimensional data, community detection and incorporating network information in hypothesis testing problems and regression analysis. The models and methods are illustrated with examples drawn from diverse fields, including biology, social sciences and text mining.

Post lunch – Session A — Xerox Auditorium, Building 9

Tutorial on the R Statistical Language and Programming Environment

Session chair: Ms Katie Evans

2:30-2:50pm **Statistical Analysis in R – An Introduction**

Christopher Claeys, Rochester, NY
eMail: cmclaeys@gmail.com

R is quickly becoming the statistical package of choice for businesses and academia alike, yet it has one of the steepest learning curves for beginners. This presentation will be pragmatically oriented and focus on giving an overview of the most essential functions of R for performing basic statistical analysis. Further, the presentation will culminate with an example of performing a linear regression analysis, evaluating the output, and discussion of the process along the way.

DAT-UP 2013

Announcing DAT-UP 2013, the first Data Analytics Tournament of Upstate New York, hosted by the Rochester Chapter of the American Statistical Association and the Journal of Unified Statistical Techniques (JUST). Form your team now, and compete to win a coveted Data Analytics Champion's trophy along with some prize money.

Eligibility: Undergraduate and Master's Students from any University or College in Upstate New York. No restriction on your major.

Teams: 3-4 students along with a faculty mentor.

Dates: Register your team by June 10, 2013

Contact: Dr Ernest Fokoué, ernest.fokoue@gmail.com

Post lunch - Session B — Room 2159, Building 9

Variable Selection and Model Selection Techniques

Session chair: Mr Nicholas Korach

2:30-2:50pm **Internship Experience in Analyzing Vaccine Stability Data**

Charlene Pu

eMail: hxp2687@rit.edu

Nine-month internship experience working as a statistical analysis intern with Sanofi Pasteur, which is a leading vaccine manufacturer. Primary responsibility is to conduct Partial Least Square (PLS) analysis to quantify potential influential factors for vaccine stability and to evaluate concentration stability data in order to determine the formulation targets that meet the concentration limit on both release date and expiration date with different confidence levels. Multiple statistical software are used, including JMP and SAS.

2:55-3:15pm **Identifying Subgroups with Enhanced Response to Treatments using Random Forests**

Mr Padraic Neville, Fairport, NY and SAS, Cary, NC

eMail: Padraic.Neville@sas.com

We seek a simple yet reliable characterization of people who respond well to a treatment. Our method starts with a random forest. Tree nodes are split to maximize a differential response to treatment. Leaves with many observations in common are then clustered. The average treatment effect (ATE) in a leaf is computed with out-of-bag data. A cluster in which the distribution of ATE is sufficiently high is investigated for characterization. We believe viewing the distribution of ATE in clusters of similar subgroups before selecting a subgroup mitigates reversion of the treatment effect with future data in the subgroup.

This report is preliminary; validation is incomplete. This paper illustrates the method, explains how to evaluate splitting criteria based on their sensitivity to a signal, and illustrates how the predictions from forests can rank prospects for treatment better than published tree-based methods in promotional marketing

Rochester Institute of Technology (RIT)

Afternoon – Session A — Xerox Auditorium, Building 9

Tutorial on computational approaches to statistical theory

Session chair: Ms Sara Smacher

3:30-4:25pm

Computational Demystification of Some BasicTheoretical Statistics Results

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

I present in a workshop/lab style, some computational demonstrations of fundamental theoretical statistics results such la definition of frequentist probability, the weak law of large numbers, the strong law of large numbers, the central limit theorem, the biasedness of the MLE for the variance, Chebyshev's inequality, Chernoff Inequality, the coverage of Ward's confidence interval, the empirical view of the significance level of hypothesis test, the empirical view of the power of standard test, just to name a few. The goal of the presentation is to provide teachers with a possible way to make fundamental theoretical statistics results more intuitive

.

Afternoon – Session B — Room 2159, Building 9

Excursion into Statistics in Sports

Session chair: Dr Chulmin Kim

3:30-3:55pm

Give Me A Break: A Quantitative Analysis of Success Factors in the Association of Tennis Professionals

Nick Korach, Rochester Institute of Technology
eMail: ndk9985@rit.edu

In the ever-growing field of Sports Statistics there has been very little research done involving tennis. However, when watching a tennis match on television the broadcasters always seem to reference various "match stats." It seem as though the commentators constantly talk about several of these factors, such as number of aces, number of break point opportunities, and first serve percentage. In this presentation, we will examine 18 of these "match stats" accumulated for the top 100 male singles players over the past five years. Using both supervised and unsupervised learning, we will determine which of these factors are most significant in defining success of a male singles player.

4:00-4:25pm

Predicting success in the National Football League - An in-depth look at the factors that differentiate the winning teams from the losing teams

Ben Rollins, Rochester Institute of Technology
eMail: bjr9780@rit.edu

Almost all American Football experts seem to agree, that of all the statistical measures used to track the performance of football teams, third down conversion percentage stands out as the one with the highest predictive power, especially at professional levels like the National Football League. However, existing literature on American football statistics provides very little in the way of a deeper look into the tactical reasons why this single down in football seems to discriminate so well between successful and unsuccessful teams. In this paper, we propose and explore in great details a variety of statistical models and techniques aimed at capturing the inner workings of third down conversion, and we suggest ideas on how coaches may potentially tweak their play calling to improve their third down conversion percentage and thereby achieve greater success. We also hint a simulation studies geared towards demonstrating how a team can tactically improve their third down conversion percentage by implementing the findings of our statistical analyses.