SESSION 1A

Biostatistical Applications

9:30-9:50  The statistics of suicides

Jennifer Bready
Division of Math and Information Technology, Mount Saint Mary College
E-mail: Jen.Bready@msmc.edu

This presentation will discuss trends in suicides from data compiled in Dutchess County, NY from 2003 to 2015. These data were based on reporting and examination standards recommended by the National Association of Medical Examiners and Department of Justice death investigation standards. Over 350 cases were analyzed in terms of the decedent’s gender, age, and race, the presence of a suicide note, method of suicide, weapon used, and toxicology results. The presenter will discuss the statistical analysis of the suicide data, make comparisons to national data published by the CDC, and will share why this statistical analysis is important to the county (and other counties). Results from this study agree with other reports showing significant differences for method of suicide in terms of both age and gender, with men more likely to choose gunshot and women more likely to choose overdose. Furthermore, older persons were more likely to commit suicide by gunshot, and the younger persons were more likely to commit suicide by overdose. However, results in other areas, such as method of suicide and presence of a suicide note, differ from national statistics.

9:55-10:15  Order-of-Addition Experiments

Joseph Voelkel
School of Mathematical Sciences, Rochester Institute of Technology
E-mail: joseph.voelkel@rit.edu

The order in which components are added in a chemical batch, film, food products, or a study of protein transport may be a primary consideration in an experiment. We review work that has been used for such order-of-addition (OofA) experiments, and propose extensions to them. We define a reference standard of OofA experiments by extending the idea of orthogonal arrays. For strength 2 designs, upon which we focus most of our attention, we find that OofA orthogonal arrays require $N = 0 \mod 6$ runs. We consider a $\chi^2$ criterion for the balance of the array, and also a Hamming-distance criterion among the rows of the design. We find empirically that D-optimal designs (using a particular set of columns in the model matrix) usually perform very well under each of our two criteria. For these reasons, as well as its general availability, we recommend this method for design construction. We then extend these optimal OofA designs to incorporate standard process variables so that, for example, temperature or mixing speeds may be included as well. Our methods also take into account natural restrictions that the experimenter may have, such as requiring that one component is always added before another. Our main emphasis is on the design of OofA experiments, but we also provide some suggestions and examples for analysis.

SESSION 1B

Advances in Clustering Methodology Motivated by Mixed Data Type Applications

9:30-9:45  Hitting the wall: mixture models of long distance running strategies

Joseph Pane
Department of Statistics, Carnegie Mellon University

§
The International Association of Ultrarunners holds an annual 24 Hour World Championship where the winner is the entrant who completes the longest distance during the 24 hours. As in all races, there are different strategies associated with the runners, some by choice (e.g., planned in advance) and some by circumstance (e.g., injury, unexpected weather). Using the number of completed laps per hour for each entrant, our goal is to determine the number and type of different strategies among elite ultrarunners. After a brief overview of mixture models and latent class analysis, we present a mixture model approach using variables of mixed type (continuous, ordinal, and nominal), extending the work of McParland and Gormley. In general, the estimation capability of this complex model limits the model types and numbers of clusters available. More specifically, we found that well-separated clusters with respect to nominal variables can cause difficulties in the related Monte Carlo approximations used to calculate expected probabilities in the E-M algorithm. We address that problem and illustrate some of the advantages and disadvantages of the approach using both simulated runners and data from the 2012 Ultrarunner 24 Hour World Championship held in Katowice, Poland.

9:45-10:00 Prediction via clusters of CPT codes for improving surgical outcomes §

Elizabeth Lorenzi
Department of Statistical Science, Duke University
E-mail: ecl21@stat.duke.edu

We develop a novel algorithm, Predictive Hierarchical Clustering (PHC), for agglomerative hierarchical clustering of current procedural terminology (CPT) codes, with the goal of finding clusters that improve the performance of a sparse logistic regression model for predicting surgical outcomes. The clustering scheme mimics traditional Hierarchical Clustering; however, our merge criterion is not based on a distance function and does not initialize with $n$ clusters. Our predictive hierarchical clustering aims to cluster subgroups, not individual observations, found within our data, such that the clusters discovered result in an improved performance of a classification model. Therefore, merges are chosen based on which pairings of the subgroups result in the largest improvement in prediction, as measured by the area under an ROC curve. The motivation is to predict patient-specific surgical outcomes using data from ACS NSQIP (American College of Surgeon’s National Surgical Quality Improvement Program). An important predictor of surgical outcomes is the actual surgical procedure performed as described by a CPT code. We use PHC to cluster these subgroups together in a way that enables us to better predict patient-specific outcomes, instead of currently used clinically decided clusters. We present two different configurations of our algorithm, one incorporating the clusters of CPT codes as random slopes and the second as random intercepts in the classification model.

10:00-10:15 Clustering text data for record linkage

Samuel Ventura
Department of Statistics, Carnegie Mellon University
E-mail: sventura@stat.cmu.edu

Record linkage, or the process of linking records corresponding to unique entities within and/or across data sources, is an increasingly important problem in today’s data-rich world. Due to issues like typographical errors, name variation, and repetition of common names, linking records of unique entities within and across large data sources can be a difficult task, in terms of both accuracy and computational feasibility. We frame record linkage as a clustering problem, where the objects to be clustered are the text records in the data source(s), and the clusters are the unique entities to which the text records correspond. We use the following three-step approach for record linkage: First, records are partitioned into blocks of loosely similar records to reduce the comparison space and ensure computational feasibility. We propose a sequential blocking approach that iterates through a nested set of decreasingly strict blocking criteria to reduce the comparison space more efficiently, with clusters of records being identified at each iteration. Second, we adopt an ensemble supervised learning approach to estimate the probability that a pair of records matches. We propose a new adaptive prediction approach for classifier ensembles (specifically, random forests) that extracts and incorporates summary statistic information from the distribution of estimated probabilities. Third, after transforming our estimated pairwise probabilities of matching to pairwise dissimilarities, we use hierarchical clustering to link matching records. We apply these approaches to two labeled record linkage datasets: a set of labeled inventors from the United States Patent and Trademark Office database and a compilation of lists of death records from the Syrian Civil War conflict.
Undergraduate/Graduate Statistics Education

9:30-9:50 Psychological statistics and the Stockholm syndrome

Susan Mason, Sarah Battaglia, and Sarah Ribble
Department of Psychology, Niagara University
E-mail: sem@niagara.edu

For many psychology majors, the required course in statistics is seen as a necessary evil. The students fear a heavy workload, challenges beyond their skill levels, public embarrassment, failing grades, even the end of their dreams of graduate school and a career. Moderate levels of anxiety can motivate students to work hard and can, therefore, contribute to successful performance. However, when the fear of failure is so great that it paralyzes a student, it can be a self-fulfilling prophecy. Recognizing the relationship between anxiety and success in a statistics course, we developed a course structure and support system to reduce student anxiety levels and optimize student performance. Critical factors in our program include small class sizes; clear expectations; frequent practice in class, on homework, and on tests; regular, specific feedback; and significant support from the professor and a teaching assistant. In this presentation we outline our course system, review a series of studies evaluating the system, and offer suggestions for future research and application. Why did we title our presentation “Psychological Statistics and the Stockholm Syndrome”? We chose that title because students often begin the course with high anxiety levels, like hostages expecting to be tortured. By the end of the semester, though, student course evaluations reveal that the “hostages” (students) have developed more positive feelings toward their “captors” (the professor and the teaching assistant) and about the course. Even the students who at the beginning of the semester rate their anxiety levels at 11 on a 10-point scale can find success. With that success they develop confidence, work harder and accomplish more. Their sense of personal pride turns to appreciation for those who believed in them and provided the opportunity to succeed.

9:55-10:15 Interdisciplinary professional science master’s program on data analytics between SUNY Buffalo State and SUNY Fredonia

Joaquin Carbonara and Valentin Brimkov
Department of Mathematics, SUNY Buffalo State
Renata Barneva
Department of Applied Professional Studies, SUNY Fredonia
E-mail: carbonjo@buffalostate.edu

There is a high demand for experts in data analytics. Some large universities have already opened programs in the field. Unfortunately, they are not always affordable for students, who in most cases already have a bachelor or master’s degree and are looking to augment their knowledge and skills or to start a new career. At the same time, in the current economic situation in academia it is sometimes very difficult to start a new program and hire faculty with the necessary expertise. SUNY Buffalo State and SUNY Fredonia united their efforts and course offering capabilities to propose a joint interdisciplinary program on data analytics. It will be offered as a Professional Science Master’s (PSM) program and courses will be taught in both universities as most of them will be offered in a synchronous distance education manner or on-line so that the students can take them from various locations. In addition, industrial partners will be included widely as internship mentors or project advisors. This will make the program very flexible and will ensure invaluable practical experience and university-industry partnerships. In this talk we will present the program structure, the courses it includes, and the opportunities it offers.

Biostatistical Methods

9:30-9:50 Detecting changes in matched case count data three ways: analysis of the impact of long-acting injectable antipsychotics on medical services in a Texas Medicaid population
We describe three known modeling approaches to a matched case study. Claims data from approximately 3000 patients enrolled in Texas Medicaid were analyzed to detect any changes in a number of outcomes after initiation of a long-acting injectable antipsychotic. Statistically, outcomes are counts of events (hospitalizations, office visits, etc.), before and after a regime change for each patient. The task is to assess any differences in rates of events. Data were analyzed three ways, each with varying levels of sophistication. Because outcomes were matched, count differences for each patient can be used as observations. We compare the results of three modeling efforts: a linear longitudinal model, a simple bootstrap of count differences and a Bayesian fitting of a zero-inflated difference of Poisson random variables model. Each approach has different modeling assumptions, diagnostics and extensions. Nevertheless, conclusions resulting from each method were similar for this study. The linear longitudinal model when fitted in a Bayesian paradigm leads to greater modeling flexibility for future investigation.

9:55-10:15 Bayesian approaches to missing data with known bounds: dental amalgams and the Seychelles Child Development Study

Chang Liu and Sally Thurston
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
E-mail: salvatorecliu@gmail.com

Dental amalgams are approximately 50% metallic mercury which releases mercury vapor (Hg\textsuperscript{0}) throughout lifetime after restoration. Chronic exposure to elevated levels of Hg\textsuperscript{0} can result in neurotoxicity. However, the risk of prenatal exposure to Hg\textsuperscript{0} on the neurodevelopment of the children is not known. In the Seychelles Child Development Study (SCDS), we are interested in evaluating such potential adverse health effects from a Bayesian perspective. As a surrogate for the levels of Hg\textsuperscript{0} exposure, the numbers of dental amalgams are often unknown and only the lower and upper bounds are available. Naïve linear regression models may not lead to accurate results. We propose to use a zero-inflated negative binomial distribution to model the missing data and incorporate the bound information in posterior inference. An iteratively reweighted least squares Metropolis-Hastings algorithm and latent variable approach will be applied to improve convergence.

SESSION 1E

Machine Learning

9:30-9:50 An introduction to ensemble methods for machine learning

Kenneth Tyler Wilcox
School of Mathematical Sciences, Rochester Institute of Technology
E-mail: ktw5691@rit.edu

While single models can clearly and effectively explain data, particularly when inference and interpretation is a priority, aggregation of multiple models often provides tremendous gains in inference and prediction. Aggregated models, commonly known as ensembles, are a powerful addition to a statistician’s and data scientist’s “toolbox.” I provide an overview of several current benchmark ensemble algorithms in contemporary machine learning. My focus includes general bagging, random subspace learning, and boosting algorithms; the powerful random forest algorithm proposed by Breiman in 2001 provides a focal point for this discussion. Such ensembles commonly deliver unmatched predictive accuracy without overfitting training data; we explore the ways in which ensembles balance the trade-off between estimation bias and variance. Furthermore, we emphasize the use of ensembles in nonparametric function estimation. An intuition for their use, strengths, and weaknesses is developed through
exploration of data sets using implementations of ensemble algorithms implemented available in the R programming language. Finally, extension of ensembles to survival analysis and future research is explored.

**9:55-10:15 Generalization of training error bounds to test error bounds of the boosting algorithm**

Paige Houston and Ernest Fokoué  
School of Mathematical Sciences, Rochester Institute of Technology  
E-mail: pjh1123@rit.edu

Boosting is a unique machine learning method which uses a base model that is only slightly better than random guessing to improve accuracy of prediction or classification. Accuracy improves because each misclassified point helps the next model essentially learn from its mistakes, making each iteration after that more accurate. Because of this, the training error becomes strictly decreasing as the number of rounds increases. Many researchers have explained and proved the training error bound theorem regarding the classification method of boosting. This unique trait of the boosting algorithm intrigued us to see if there is some sort of generalization to the test error as well as a possible test error bound. I will discuss how the training error bound was discovered and how the theorem was constructed. We will then use R to boost several datasets to analyze the relationship between the empirical training error and test error. Finally, we will project the training error bound to create a potential test error bound.

**SESSION 1F**

Geological Applications / Image Generation

**9:30-9:50 A new approach to quantifying stratigraphic resolution: measurements of accuracy in geological sequences**

H. David Sheets  
Department of Physics, Canisius College  
E-mail: sheets@canisius.edu

Geochronology is the process of forming a timeline of events in earth history, based on individual collections of fossils, radio-isotope dates, environmental measurements, paleomagnetic measurements, and a range of other geological information. Classically, geochronology has been calculated using graphic correlation, an iterative pairwise process. Since the early 1980s, automated approaches to geochronology have appeared, mostly based on optimality criteria. We report here on the use of a second generation automated geochronology tool, Horizon Annealing (HA), which provides a method to order all horizons in a chronostratigraphic data set. International time standards are based on Global Stratotype Section and Point (GSSP) levels, which may be placed within an HA composite succession and be precisely correlated with levels in all other sections in the composite that spans the same interval. We present two approaches to the quantitative assessment of the uncertainty or error in the placement of horizons within the composite section: a permutation method (jackknife analysis) and a sensitivity analysis (the relaxed fit curve). These approaches provide an objective method for assessing the relative strengths of proposed GSSP candidates, in which we prefer those sections and horizons that are the most precisely controlled in the HA composite. Integration of biostratigraphic, chemostratigraphic and lithostratigraphic marker horizons into the HA process markedly improves levels of constraint within the composite compared to biostratigraphic data alone.

**9:55-10:15 Image generation in the era of deep architectures**

Ifeoma Nwogu  
Department of Computer Science and Engineering, SUNY at Buffalo  
E-mail: inwogu@buffalo.edu

In recent years, deep learning methods have proven useful for handling complex, high-dimensional data such as images, videos, audio, etc., by taking advantage of shared and distributed representations while learning. With the successes of these new architectures, the age-old computer vision problem of modeling and generating natural images and videos is revisited again. There have been many recent successes in this area although the resulting images are still of lower resolution and quality. To this end, by building on the recent successes of several deep learning...
techniques that possess one or more characteristics of an ideal generative model, we propose an architecture that readily facilitates the generation of independent, highly naturalistic, high-resolution image samples. This architecture also allows for efficient inference when data are presented to it. We demonstrate its efficacy by performing quantitative tests on the standard benchmark datasets, and qualitative tests on large, diverse complex datasets (over two million high-resolution 256 x 256 images).

SESSION 2A

Strategies for Analyzing Multiple Outcomes in Biostatistics

10:25-10:45 Clustering multiple outcomes via the Dirichlet process prior
Amy LaLonde and Tanzy Love
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
E-mail: Amy_LaLonde@urmc.rochester.edu

Environmental exposure effects on human development can be small and difficult to detect from observational data. In the Seychelles Child Development Study, researchers examined the effect of prenatal methylmercury exposure on a battery of 20 tests measuring aspects of child development. Thurston et al. posed a multiple outcomes model with outcomes nested in domains to allow covariate and exposure effects to differ across domains, and multi-level random effects to fully capture correlations among outcomes in the same domain and across domains. Xiao et al. extended the model by allowing partial membership of outcomes in domains. Our model assumes single domain assignments for outcomes, yet clusters the outcomes into domains with a Dirichlet process prior and a Bayesian MCMC approach. Results show more pronounced exposure effects in one of the domains, distinct domain-specific covariate effects, and sensible domain assignments.

10:50-11:10 Global tests for multiple outcomes in randomized trials
Donald Hebert
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
E-mail: Donald_Hebert@urmc.rochester.edu

Global tests are highly useful for randomized trials in diseases with multiple outcome measures of equal importance. Such methods play a pivotal role in assessing an overall treatment effect and are particularly powerful in the case where the treatment effects on individual outcomes are consistent, i.e., in the same direction. Much attention has been given to this problem when the outcomes are assumed to follow a multivariate normal distribution. O'Brien's tests based on ordinary and generalized least squares are very powerful when the individual effects are similar in magnitude. When the effects are dissimilar, these procedures lack power compared to other tests such as the approximate likelihood ratio test. This talk will present a novel class of global tests based on procedures for combining p-values derived from orthogonalized test statistics. Simulation studies demonstrate that these tests can provide high power that is stable across different alternatives in the positive orthant. Outcome-specific inference and extensions to outcomes of mixed type will also be discussed.

SESSION 2B

New Progress in High-Dimensional Data Analysis with Dimension Reduction

10:25-10:40 Consistency and convergence rate for the nearest subspace classifier
Yi Wang
Department of Mathematics, Syracuse University
E-mail: ywang392@syr.edu

The nearest subspace classifier (NSS) solves the classification problem that is, given labeled samples, to find a rule that assigns an arbitrary sample to a class. More precisely, NSS obtains an estimation of the underlying subspace
within each class and assigns a data point to the class that corresponds to its nearest subspace. I will discuss how well NSS can be generalized to new samples. Conditions under which NSS is strongly consistent and has a rate of convergence $O\left(\sqrt{n}\right)$ will be given. Some simulations will be demonstrated to verify the theoretical results.

10:40-10:55  A new approach to sparse sufficient dimension reduction with applications to high-dimensional data analysis

Wei Qian  
School of Mathematical Sciences, Rochester Institute of Technology  
E-mail: wxqsm@rit.edu

Sufficient dimension reduction (SDR) consists of a class of powerful regression analysis tools that find the minimal linear reduction subspace (known as the central subspace) to allow efficient study of the predictor relationship with the response. We will survey the ideas behind SDR methods and propose a new approach to sparse SDR methods to allow simultaneous predictor screening. We propose computationally fast algorithms and present consistency results in both central subspace estimation and predictor selection. Effective applications of the proposed approach in high-dimensional data analysis are established in both theory and numerical examples.

10:55-11:10  Tensor sliced inverse regression with application to neuroimaging data analysis

Shanshan Ding  
Department of Applied Economics and Statistics, University of Delaware  
E-mail: sding@udel.edu

Sliced inverse regression (SIR) is a widely used nonparametric method for supervised dimension reduction. Conventional SIR mainly tackles simple data structure but is inappropriate for data with array (tensor)-valued predictors. Such data are commonly encountered in modern biomedical imaging and social network areas. For these complex data, dimension reduction is generally demanding to extract useful information from abundant measurements. We exploit the higher-order structure and propose a new method known as tensor SIR, which reduces a tensor-valued predictor’s multiple dimensions simultaneously via tensor decomposition. The proposed method provides fast and efficient estimation and avoids the necessity of high-dimensional covariance matrix inversion. We further show its asymptotic behavior and numerical performance in a neuroimaging data application.

SESSION 2C  
Room XXX  

Statistical Methods and Computational Tools for Modeling the Spread of Infectious Diseases

10:25-10:45  Statistical and computational tools for generating synthetic ecosystems

Lee Richardson  
Department of Statistics, Carnegie Mellon University  
E-mail: leerich@cmu.edu

Modern approaches for modeling the spread of infectious diseases rely on the existence of detailed, person-level information about populations in specific geographic regions. Because government agencies such as the U.S. Census Bureau are prevented by law from releasing personally identifiable information about individual persons, researchers often use synthetic populations instead. Synthetic populations avert the issue of releasing private information about individuals by instead releasing detailed information about synthetic people. Using public-use microdata samples, which provide detailed information on small, representative subsets of people, we estimate the characteristics of the broader population. Then, using these estimates in conjunction with population counts in specific geographic regions, we sample a new, synthetic population. For example, in our synthetic population, each sampled person has characteristics such as age, gender, income, etc. Ideally, the marginal and joint distributions of these characteristics align with those of the actual population within each geographic region. We discuss the statistical and computational challenges of generating synthetic populations for geographic regions across the entire
First, we introduce the different approaches for sampling households, locations, and people, including a bootstrap-like approach, iterative proportional fitting, and joint density estimation. Next, we develop an open-source R package, SPEW, that provides a general framework for using these techniques to generate synthetic populations for any geographic region. Specifically, SPEW can incorporate a wide variety of input data sources, sampling methodologies, and algorithms for assigning individuals to elements of their environment (e.g., schools and workplaces). After a comprehensive data collection effort, our primary computational challenges are in standardizing across heterogeneous data sources, parallelizing our algorithm for use on a high-performance computing cluster from the Pittsburgh Supercomputing Center, and assessing the optimal subpopulation size for processing efficiency.

10:50-11:10 **An overview of statistical modeling of infectious diseases**

Shannon Gallagher  
Department of Statistics, Carnegie Mellon University  
E-mail: sgallagh@andrew.cmu.edu

Modeling the spread of infectious diseases is an important problem in modern public health. We overview several statistical approaches for predicting, simulating, and analyzing the spread of various viruses and diseases. First, we produce a new model for predicting influenza incidence in the United States. The new model is an extension on an Empirical Bayes model proposed by Brooks et al. Our model, Posterior Biasing, focuses on predicting the primary target of the peak incidence of influenza in 10 different regions in the United States. Posterior Biasing performs twice as well as those generated by Empirical Bayes in predicting the peak week of influenza. Moreover, we produce probabilistic forecasts through our posterior distribution and compare our Bayesian credible intervals with the frequentist coverage, which empirically seem to have a one-to-one correspondence. Next, we discuss statistical methods for modeling the spread of recent infectious diseases such as Zika, Ebola, Chikungunya, and the “next” disease. We detail how synthetic populations and ecosystems can be used in agent based modeling. We also provide an overview of susceptible-infected-recovered (SIR) models, for which we use linear regression to directly infer the reproduction rate. Finally, we build an interactive exploratory data analysis tool for visualizing disease incidence rates in the United States since 1888.

**SESSION 2D**  
**Analytics in Sports**

10:25-10:45 **Statistics and data analytics in sports management**

Renata Barneva  
Department of Applied Professional Studies, SUNY Fredonia  
Valentin Brimkov  
Department of Mathematics, SUNY Buffalo State  
Patrick Hung and Kamen Kanev  
Faculty of Business and Information Technology, University of Ontario Institute of Technology  
E-mail: barneva@fredonia.edu

Sports management is a multi-billion industry that is heavily influenced by modern technologies and software advancements. While novel sports devices and advanced equipment are indispensable for top-level performers in competitions, ordinary people also take advantage of them to exercise and keep in shape. On the one hand, body motions, heartbeat, blood pressure, and other sports-related information can easily be collected by modern smart phones, smart watches, and other wearable devices and stored in the cloud as Big Data. Software advancements and data analytics, on the other hand, provide the means for targeted statistics and analysis of such data that are poised to change the way sports are played, enjoyed, and sponsored. Quantitative and qualitative data about team compositions and individual players, scores and penalties, game tournaments and ticket sales, and much more are already available in professional sports data collections. Data analytics performed on such data could be used, for example by (1) managers for optimal composition of teams and attracting the best players, (2) coaches in training sessions and for game and playground arrangements, (3) sponsors for identifying sports investment options and optimizing the returns, (4) bookmakers for calculating the odds and thus offering better bets, and others. Statistics and analytics applied to the even bigger volumes of exercise data collected in gyms are of great value to ordinary
people providing for selective feedback and intelligent advice on continuing sports activities. After a brief survey, we will present examples of our research on data analytics in sports – more specifically in fencing and combining physical exercises with kanji learning – and how it can be used to improve performance and retention.

10:50-11:10 Predictive analytics tools for identifying the keys to victory in professional tennis

Shruti Jauhari and Aniket Morankar  
Department of Computer Science, Rochester Institute of Technology  
Ernest Fokoué  
School of Mathematical Sciences, Rochester Institute of Technology  
E-mail: asm6887@rit.edu

The history of the beautiful game of tennis is filled with the names of exceptional players who dominated their era with unusually long winning streaks, documented spectacular performances, large numbers of grand slam victories and marquee tournament trophies, just to name a few. In recent years, in both the ladies and the gentlemen versions of the game, we have witnessed the emergence of great names and intense rivalries inspired by the exceptional prowess of legendary players. As one watches the dominance of those legendary performers, one is tempted to wonder how they manage to vanquish their opponents repeatedly and sometimes with relatively great ease. What are the factors or elements of their performance that consistently give them the advantage/edge? Most experts, pundits and analysts have argued that serving powerfully and consistently is the master key to tennis dominance. Others believe that the athletic ability to dominate court coverage and consistently return the ball is the holy grail of tennis success. Still others believe that all else being equal, focus is the key to consistent/repeated success in tennis. In this project, we collected statistical data from four Grand Slam tournaments played in 2013 and 2015 as well as data from some ATP and WTA tournaments to find out which factors/aspects of the game tend to associate most strongly with consistent and repeated success in the game of tennis. We use techniques and methods of high level programming language, tools of modern data mining and machine learning such as classification and regression trees, logistic regression, boosting, Microsoft C#, HTML agility pack, Gembox spreadsheet and firebug to perform data collection, variable selection and predictive analytics.

SESSION 2E  
Room XXX

Statistics in the Evaluation of Education

10:25-10:45 An analysis of covariance challenge to the Educational Testing Service

Richard Escobales  
Department of Mathematics and Statistics, Canisius College  
Ronald Rothenberg  
Department of Mathematics, CUNY, Queens College  
E-mail: escobalr@canisius.edu

In 1979, the Educational Testing Service (ETS) issued a preliminary report comparing the mathematical performance of two groups of New York students. One group used the traditional Regents sequence for high school mathematics; the second used the so-called Integrated Program. Although the traditional students had a higher percentage of correct answers to 42 of the 50 mathematics questions, the ETS concluded that a student who participates in Integrated Program would have no practical disadvantage over a student of similar (verbal) ability who participated in the traditional Regents Program. We use analysis of covariance to challenge this conclusion.

10:50-11:10 Model, model, my dear model! Tell me who the most effective is

Yusuf K. Bilgic  
Department of Mathematics, SUNY Geneseo  
E-mail: bilgic@geneseo.edu

Data analysis in educational and social science research often addresses questions related to the examination of effects and the study of the factors that explain differences. Until the 1970s, ordinary linear modeling was seen a
magic box to help to reveal statistical patterns; however, it was inappropriate for correlated and nested structured
data, and “better” models were introduced later. Currently, as a déjà vu, the effectiveness of school teachers, 
principals and educational interventions has been analyzed using “even better” models such as value-added models 
(VAMs) adopted by many districts/states for decisions. Data analyses and decisions with statistical models are great 
scientific powers; however, they lack the ability to handle the big picture and hinder some realities. In this talk, I 
will introduce some famous statistical models in general being used for assessment and evaluation purposes: VAMs, 
value-added research centers (VARCs), student growth percentiles (SGPs), Growth, and Multilevel/Hierarchical 
models along with their limitations.

SESSION 2F

Data Science / Computing

10:25-10:45  An example of four data science curveballs §

Tamal Biswas and Kenneth Regan
Department of Computer Science and Engineering, SUNY at Buffalo
E-mail: tamaltan@buffalo.edu

Data science and computational experiments involve assumptions of design and need for controls as in any other 
science. The assumptions may be buffeted by unsuspected phenomena in the algorithms that compile the data. We 
present four such phenomena that have arisen in the study of human decision making at chess via computer analysis 
of millions of chess positions from games by players of all skill levels. That programs on commodity hardware now 
far outstrip all human players two decades on from the Kasparov-Deep Blue matches has opened new vistas for 
authoritatively judging and predicting decision behaviors, but awareness of the programs’ quirks is needed: (1) 
Moves that are given identical quality scores by the programs might be treated alike in models, but human players 
prefer the one the program lists first by a whopping and consistent 58%-42% margin. How can this be, since if the 
program randomly switched the listing order its output would be just as valid – the moves being peas-in-a-pod from 
their view? It seems like a wildly successful ESP test. The answer traces to the programs’ use of stable sorting 
algorithms and the sub-surface human impact of depth in calculations. (2) The programs can be asked to evaluate 
yea-many secondary moves fully, as opposed to the normal playing mode of stopping once the program is satisfied 
they are inferior to its best move. One might expect this not to affect the human frequency of having played said 
best move, but it does by 1-2%, regardless of the skill of the human player. (3) Chess programs use hash tables that 
fill in seconds as millions of positions are searched; happily, collisions are relatively harmless. Clearing the hash 
before each move should not make a difference – and doesn’t for the strongest players – but for all other players it 
artificially depresses the best-move agreement by 0.5%. (4) Human players are observed to make twice as many big 
mistakes when the computer judges them behind by a minuscule amount – under a tenth of a pawn’s value – as 
when they are ahead by the same tiny amount. For the strongest players, the gap expands to six times, completely 
destroying a natural “continuity” assumption in the model. Are they psyched out? Can this possibly be a real 
phenomenon? We argue it is a chaos-theory phenomenon in chess itself. It would be easy to overlook these 
phenomena when developing a model, but missing any one throws it off considerably. The issues will be framed in 
student-friendly discussion of statistical modeling and what your data mean. The role of stable sorting carries 
separate interest.

10:50-11:10  Initializing R objects using data frames: tips, tricks, and some helpful code to get you 
started

Donald Harrington
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
E-mail: Donald_Harrington@urmc.rochester.edu

Object initialization (program variables and properties, regression formulas, etc.) is often performed as needed or 
“in-line” with program flow in R and many other programing languages. For analyses requiring many lines of code, 
assignments can be deeply embedded and difficult to find. In the case of complicated analyses requiring many 
different models and formulas of varying type and method, code can become bloated and formulas difficult to 
manage based on the assignment of each required model. There is an alternative approach to initializing analysis 
parameters in R that involves storing program specific properties data in delimited files, reading those files into R, 
and then processing those files to initialize the objects. For example, regression formulas can be stored in one data
file and program options (say the alpha value or file locations) stored in another. The creation of these initialization
files can be done with a text editor, Excel, or other software that creates delimited files. The initialization files then
provide the statistician with documentation and a single location to find assignments. This session will provide a
brief introduction to the use of this alternative approach, as well as some tips, techniques, and traps that this
statistician has learned by examining a regression example with many formulas and methods.

SESSION 3A

Undergraduate Statistics Education

11:20-11:45 Undergraduate statistics: How do students get there? What happens when they leave?
(and everything in between . . .)

Rebecca Nugent and Paige Houser
Department of Statistics, Carnegie Mellon University
E-mail: rnugent@stat.cmu.edu

Carnegie Mellon Undergraduate Statistics is in a period of unparalleled growth, ballooning from 10-12 majors to
about 250 majors in several different Statistics tracks or concentrations. The demand for modern statistics classes is
not constrained to majors only; our undergraduate program now teaches about 1500 students each semester (20-25% of
CMU’s undergraduate population) from mathematics, computer science, business, psychology, neuroscience, etc. As a result, we have completely revamped our advising framework in order to continue providing (what we
hope is) a high quality, personalized experience. We have also redesigned our curriculum to be more flexible in
addressing the often competing needs of students going into industry and those going to graduate school. Some new
additions include data science competitions, dedicated time and effort to scientific writing, and the replacement of
textbook problems with interdisciplinary research projects. This panel presentation and subsequent discussion will
give an overview of our recent program changes including lessons learned and future plans. We will pay special
attention to the freshmen and sophomore years where students are transitioning to a university setting and then
transitioning to a Statistics major. In particular, we will give a snapshot of the struggles that students face and how
they might better prepare for undergraduate statistics. We also summarize the variety of post-graduate choices our
students make. Several CMU undergraduate Statistics alumni will also provide their personal perspectives.

11:45-12:05 Panel discussion

SESSION 3B

Health Policy Statistics

11:20-11:35 Application of two stage residual inclusion (2SRI) model in testing the impact of length of
stay on short-term readmission rate

Xueya Cai
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
E-mail: Xueya_Cai@urmc.rochester.edu

The two stage least squares (2SLS) method has been applied in health policy studies to control for endogeneity of
regression due to omitted variable bias. The two stage predictor substitution (2SPS) method is the extension of
2SLS to nonlinear models. In the 2SPS model, the predicted values of the endogenous variable obtained from the
first stage model are entered into the second stage model. The 2SPS method, however, is not statistically consistent
in all non-linear cases. The two stage residual inclusion (2SRI) method, which includes in the second stage model
the residuals obtained from the first stage model as well as the endogenous variable, is shown to be consistent in a
broader class of non-linear models. We applied the 2SRI method to test the impact of length of stay (LOS) after
coronary artery bypass graft (CABG) surgery on short-term readmission rate.
11:35-11:50 Health Services and Outcomes Research Methodology (HSORM): an international peer-reviewed journal for statistics in health policy

Yue Li
Department of Public Health Sciences, University of Rochester Medical Center
E-mail: Yue_Li@urmc.rochester.edu

Health Services and Outcomes Research Methodology (HSORM) is an international, peer-reviewed journal devoted to methods for the study of the utilization, quality, cost and outcomes of health care. Reflecting the multidisciplinary nature of the field, HSORM addresses the needs of interlocking communities: methodologists in statistics, econometrics, social and behavioral sciences; designers and analysts of health policy and health services research projects; and health care providers and policy makers who need to properly understand and evaluate the results of published research. The journal strives to enhance the level of methodologic rigor in health services and outcomes research, and contributes to the development of methodologic standards in the field. The journal publishes: research papers on quantitative, qualitative, and mixed methods; case studies describing applications of quantitative and qualitative methodology in health services and outcomes research; review articles synthesizing and popularizing methodologic developments; tutorials; articles on computational issues and software reviews; book reviews; and notices.

11:50-12:05 Jackknife empirical likelihood inference for inequality and poverty measures

Dongliang Wang
Department of Public Health and Preventive Medicine, Upstate Medical University
E-mail: wangd@upstate.edu

The majority of existing statistical inference procedures for inequality and poverty measures have been developed utilizing either asymptotic theory of U-statistics or resampling methods, such as the jackknife and the bootstrap. In this study, we develop jackknife empirical likelihood methods with a particular focus on the Gini index and the Sen poverty index, the latter of which is essentially a non-linear functional of the headcount ratio, income gap ratio and the Gini index of the poor. We derive the empirical likelihood over the jackknife pseudo-samples and establish the Wilks theorem, which allows us to develop confidence intervals calibrated by either the chi-square distribution or the bootstrap. Extensive simulation studies suggest the advantages of the proposed methods in terms of coverage probabilities and average interval widths.

SESSION 3C          Room XXX

Statistical Applications in Physics

11:20-11:40 Investigation of statistical and non-statistical fluctuations observed in relativistic heavy-ion collisions at AGS and CERN energies

Gurmukh Singh
Department of Computer and Information Sciences, SUNY Fredonia
Provash Mali and Amitabha Mukhopadhyay
Department of Physics, North Bengal University, India
E-mail: singh@fredonia.edu

The main motivation of the present investigation is to study statistical and non-statistical density fluctuations produced in the $^{28}$Si+$^{32}$S induced Ag(Br) collisions at 14.5A GeV/200A GeV from two experiments conducted at the Alternating Gradient Synchrotron (AGS) of the Brookhaven National Laboratory (BNL), Long Island, USA and the Super Proton Synchrotron (SPS) at European Center for Nuclear Research (CERN), Geneva, Switzerland. We shall discuss the details of both categories of single particle density fluctuations produced in central collisions by employing the basic technique of Multi-Fractal Detrended Moving Average (MFDMA). Multifractal parameters obtained from the big data analysis are systematically compared with the Ultra-relativistic Quantum Molecular Dynamics (UrQMD) model simulation. It is observed that the single particle density functions in both experiments are multifractal in nature. Furthermore, the degree of multifractality in the simulated-event samples is almost equal
to that in the corresponding empirical data. The results of this analysis differ significantly from those obtained from other conventional techniques of multifractal analysis formerly used for the same sets of big data.

11:45-12:05 **Analysis of big data at the Jefferson Lab**

Michael Wood  
Department of Physics, Canisius College  
E-mail: wood5@canisius.edu

The Thomas Jefferson National Accelerator Facility, or Jefferson Lab, is a cutting-edge nuclear and particle physics laboratory with the mission to understand the internal structure of the atomic nucleus. Located in Newport News, VA and operated by the U.S. Department of Energy, the lab employs a continuous electron beam in the energy range of 1 to 12 GeV. Jefferson Lab employs hundreds of scientists and engineers and is used by over a thousand scientists, both national and international. The many projects at the lab include mapping the internal structure of protons and neutrons, searching for exotic particles containing extra gluons, studying the duality of the nucleus as either a collection of protons/neutrons or a collection of quarks, and quantifying the process of a free quark forming into a stable hadron. The standard experimental procedure is to scatter the electron off of the nucleus, collect the reaction products in large detector systems, and reconstruct the reaction. A typical experiment runs continuously for 4 to 6 months and accumulates billions of reactions. This translates to terabytes of storage. This talk will describe how these big data sets are analyzed as well as how the particle information (mass, charge, energy, momentum, etc.) is reconstructed.

**SESSION 3D**  

**Room XXX**  

**Biostatistical Modeling**

11:20-11:40 **Techniques for estimating time-varying parameters in a cardiovascular disease dynamics model**

Jacob Goldberg  
Department of Mathematics, SUNY Geneseo  
E-mail: jag29@geneseo.edu

The body continuously regulates system properties to maintain blood pressure at homeostasis, which prevents fainting or light-headedness during everyday activities. This project aims at understanding how system properties are controlled during blood withdrawal by utilizing a mathematical model to predict cardiovascular dynamics. More specifically, a five-compartment model of systemic circulation is employed and validated against experimental pressure and volume data measured in the hearts of rats. Data were made available from collaborators at the University of Michigan and the Medical College of Wisconsin. The body's natural responses during blood withdrawal, such as increased heart rate and vasodilation, are represented by model parameters, and the objective of this project is to understand how these parameters change over time to predict experimental observations. Techniques used to estimate time-varying dynamics include piecewise linear splines and ensemble Kalman filtering, with the latter providing a measure of uncertainty in the parameter estimates. Future work aims at applying these techniques to patient-specific modeling and medicine to improve the quality and efficiency of treatment for underlying cardiovascular conditions.

11:45-12:05 **Prediction of event times with a parametric modeling approach in clinical trials**

Chongshu Chen  
Department of Biostatistics and Computational Biology, University of Rochester Medical Center  
E-mail: Chongshu_Chen@urmc.rochester.edu

For a clinical trial with a time-to-event endpoint, modeling and prediction of subject accrual and event time have been of interest for important practical reasons. It is common to design the interim analysis plan around the occurrence of a target number of events. The rate of occurrence of the event of interest is a significant factor that determines the timing of the analysis, upon which significant resources and strategic planning depend. The purpose
of this investigation was to demonstrate a proposed ramp-up (RU) model for prediction of landmark event times using accumulating data from a randomized trial. Currently available methods use either parametric or nonparametric models to predict the landmark event time based on accumulating information about enrollment, events, and dropouts. Bagiella and Heitjan proposed a parametric prediction model (BH model) for failure-time outcomes assuming exponential survival and Poisson enrollment, but the assumption of Poisson enrollment underestimates the true enrollment rate when there are few events at the beginning of a trial. We propose the RU model, which is an extension of the BH model with a ramp-up and steady period enrollment structure. We further implement predictive time intervals with Bayesian bootstrap simulation. The RU model is easy to apply in the early stage of a trial. The predictive interval can be used by clinical management teams to determine an appropriate time to conduct interim analyses. A limitation is that all methods reviewed in event modeling assume non-informative censoring, which cannot be tested. The predictions of landmark event times in two studies (PREVAIL and STRIVE) using the RU model were more accurate than those of the BH model. Predictions became stable after about 15% of the events had been accumulated in the PREVAIL study. The RU method can provide reliable guidance for clinical trial planning.

SESSION 3E

Statistics in Music

11:20-11:40 Can statistics make me a millionaire music producer? §

Jessica Young
School of Mathematical Sciences, Rochester Institute of Technology
E-mail: jay4588@rit.edu

If you listen to enough music, or just have an ear for it, you start to notice that some songs sound the same; they might have the same rhythm, hook, beat, etc. In particular, this is a noticeable trend among hit songs, which give the feeling that if you just splice together parts of different songs then you could make your own hit song. But is that really true? I plan on taking some hit country songs that people claim to be similar and finding out through multivariate analysis if these songs actually are measurably similar to each other and then using these same songs to randomly generate my own new “hit” song.

11:45-12:05 Automatic singer identification of popular music via singing voice separation §

Shiteng Yang
School of Mathematical Sciences, Rochester Institute of Technology
E-mail: sy2477@rit.edu

Automatic Singer Identification (Auto-SID) is a very challenging task in music information retrieval. For an Auto-SID system, there are four problems that need to be solved: vocal segment detection, separation of singing voice, feature extraction, and identification model. In this paper, a Gaussian mixture model (GMM) based on accompaniment data is used to detect vocal segment. Then robust principal component analysis is introduced to segregate the singing voice and background music. For the feature extraction method, mel-frequency cepstral coefficients and linear predictive cepstral coefficients are calculated for each frame. Finally, we introduce the GMM to model the singer's voice. The validity of the proposed Auto-SID system is confirmed via an experiment conducted on a 10-singer pop music database.

SESSION 3F

Statistical Applications in Education

11:20-11:40 Monte Carlo-based enrollment projections

Matthew Hertz
Department of Computer Science, Canisius College
Colleges and universities need predictions of continuing student enrollment levels up to five years in the future. These enrollments set budgets, determine course offering needs, and establish staffing levels. With so much relying on these enrollments, accuracy is crucial. Given future unknowns, decisions require understanding the range of potential enrollments and likelihood of each. Initially, this problem may not seem to be difficult. People often begin by thinking of the idealized situation where a student starts in the fall term of their first year, progresses through a program as a full-time student, and finally leaves with their degree. Reality is not that simple. Changes in the student population and economic climate create dramatic fluctuations in the rate at which students drop out or graduate early. Academic, financial, medical, and personal issues lead students to take semesters off and extend their enrollment beyond the normal time. Transfer students are a large, and growing, population outside the idealized model. According to a 2015 study by the National Student Clearinghouse Research Center, more than 1 out of every 3 undergraduates who started in 2008 transferred to another institution. For most schools, models will need to account for all of these possibilities in order to provide accurate results. A student’s thought processes may not be computable, but by mining enrollment records we can estimate the probability that a student will register in a given term. This presentation describes how a Monte Carlo approach using millions of simulations of a student body with these probabilities can calculate highly accurate enrollment projections and confidence intervals. Additionally, we discuss how the features used to determine students’ probabilities of returning and the populations for which this approach does not work provide insights into the factors influencing students’ choices.

11:45-12:05  Program for International Student Assessment analysis in R §

Kierra Shay  
Department of Mathematics, SUNY Geneseo  
E-mail: kjs24@geneseo.edu

Multivariate methods in statistical analysis allow for analysis of dependence and interdependence while retaining the whole data structure. A dependence technique, such as discriminant analysis and conjoint analysis, is a technique in which a set of variables are predicted or explained by other sets of variables. The goals of interdependence methods are data reduction, grouping variables together, or investigating hidden variables or components. The Program for International Student Assessment, often called PISA, allows for the opportunity to explore perennial issues in education, cross-nationally. The benchmark for success is no longer improvement by national standards alone, but increasingly in relation to the best-performing systems internationally. The study of mathematics has moved from being a requirement only for future scientists or researchers to being a fundamental component of education and applications beyond the classroom. Understanding how math scores are affected by different factors, such as gender, socio-economic status, and a variety of school dynamics, controlling for demographic structure across countries, can be achieved using the above mentioned multivariate tools. Also discussed will be troubleshooting solutions to deal with the PISA dataset and methods used for alterations, all performed in R. Exploratory analyses of the PISA data set are performed using multivariate methods.

SESSION 4

Keynote Lecture

1:25-2:35  From the classroom to the boardroom – how do we get there?

Richard De Veaux  
Department of Mathematics and Statistics, Williams College  
E-mail: rdev@williams.edu

Big data is everywhere. Companies, governments and the media can't seem to get enough of the data deluge and the tsunami of data that is about to overwhelm us and/or make us a smarter planet. But, is there any connection between the typical content of an introductory statistics course and the statistical analyses that are performed throughout the world every day? How close is this version of Statistics to real world practice?

Most courses in Statistics now start with exploratory data analysis, move on to probability and then inference. If time permits, they include models, often ending at simple regression. Unfortunately, the student is left with the impression that Statistics is a collection of tools that can help understand one variable at a time, if only the right
formula can be recalled, rather than a way of thinking about and modeling the complex world in which we live. Maybe we're teaching the course backward. We'll describe an approach that starts the course with models, making full use of students' intuition about the world and exposing them early to the power of statistical models.

SESSION 5A

Innovative Techniques to Address Missing Data in Biostatistics

3:30-3:45 Challenges in estimating model parameters in qPCR

Valeriia Sherina, Matthew McCall, and Tanzy Love
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
E-mail: Valeriia_Sherina@urmc.rochester.edu

Quantitative real-time PCR (qPCR) is one of the most widely used methods to measure gene expression. Despite extensive research in qPCR laboratory protocols, normalization and statistical analysis, little attention has been given to qPCR non-detects – those reactions failing to produce a minimum amount of signal. It has been previously shown that common methods of handling qPCR non-detects lead to biased inference. We have previously developed a method to directly model non-detects as data missing not at random and used an expectation-maximization (EM) algorithm to impute the missing data. The result is a complete data set with imputed values for non-detects using a single imputation procedure. This approach results in a sizeable reduction in bias when estimating both absolute and differential gene expression; however, it results in underestimation of variances, leading to anti-conservative inference. To address this issue, we have extended our methodology to directly estimate the model parameters via either maximum-likelihood or Bayesian estimation. We assess and compare the performance of these approaches.

3:45-4:00 Semiparametric inference concerning the geometric mean with detection limit

Bokai Wang, Changyong Feng, Hongyue Wang, and Xin Tu
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
E-mail: Bokai_Wang@urmc.rochester.edu

In biomedical research, sample geometric means are widely used to estimate and compare population geometric means among different groups. However, due to detection limits, the actual value of the measurement is not always available. A common practice is to replace missing values by small positive constants and make inference based on the imputed data. However, no work has been done to study the effect of this naïve imputation method on the inference concerning the population geometric mean. Our investigation shows that the simple imputation approach may dramatically change the outcome of the study and make the results uninterpretable, even if the detection limit is very small.

4:00-4:15 Two data analysis methods concerning missing data

Lin Ge
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
E-mail: Lin_Ge@urmc.rochester.edu

In this presentation, two multiple-criteria decision analysis (MCDA) methods, Grey Relational Analysis (GRA) and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), are proposed to deal with evaluation problems in biostatistics, based on a decision-making process with multiple targets and multiple attributes. However, during the evaluation process, missing data is one of the most key problems to consider. Therefore, we describe a rough set methodology for the analysis of incomplete data tables. A case study is used to illustrate the application of the proposed methodology.
Use of Computing in Statistics Education

3:30-3:50  
Teaching programming skills to finance students: how to design and teach a great course

Yuxing Yan  
Department of Economics and Finance, Canisius College  
E-mail: yany@canisius.edu

It has always been my belief that an ambitious finance major should master at least one computer language. This is especially true for students in quantitative finance, Master of Science in finance, business analytics or financial engineering programs. Among many good languages, R and Python are the two best ones. Based on my 6-year experience of teaching R to finance students at 3 schools, the following 7 factors are critical for designing and teaching such a course: strong motivation, a good textbook, a hands-on environment, data intensive, a challenging term project, many supporting R data sets, and an easy way to upload those R data sets.

3:55-4:15  
The biomarker challenge R package: a two stage array/validation in-class exercise

Luther Vucic, Dietrich Kuhlmann, and Daniel Gaile  
Department of Biostatistics, SUNY at Buffalo  
Jeremiah Grabowski  
School of Public Health and Health Professions, SUNY at Buffalo  
E-mail: luthervu@buffalo.edu

Our objective was to develop software to facilitate an authentic group learning activity to teach statistical design and analysis concepts relevant to a common class of genomics based biomarker experiments. A freely available R package, BiomarkerChallenge, was authored to provide the desired functionality. The package allows for the creation of simulated array and validation datasets to be analyzed during an integrative learning experience. Students are grouped together and are randomly assigned different budgets for the design and analysis of a two stage array/validation study. During the in-class challenge, array and validation data are “purchased” and then analyzed by the student groups, with each group's final results being publicly reported. The exercise concludes with the instructor providing a comprehensive overview of the results across all groups (and budgets). Software functionality includes: (1) Dropbox integration for easy transfer of results, (2) example code for student analysis, (3) instructor tools to help evaluate student progress, and (4) visualization tools to illustrate key concepts. Statistical concepts include: power, error control, multiple testing, dimension reduction, distance estimation, Fisher's exact test, and clustering techniques.

SESSION 5C

Bioinformatics

3:30-3:50  
A short survey of computational structural proteomics research using the Protein Data Bank (PDB) as the main data set

Vicente M. Reyes  
Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology  
E-mail: vmrsmbi.rit.biology@gmail.com

The Protein Data Bank (PDB) is the foremost repository of biomolecular structure data in the world, where all structural biology laboratories deposit their finished products – biomolecular (typically protein) atomic structures – usually before publication. Its home page is http://www.rcsb.org. The PDB contains mostly structures of proteins, many of which contain bound ligands, but it also has some carbohydrate, lipid and other structures of biological interest. The PDB now contains the 3D information for over 116,000 biomolecular structures, beginning with just 13 structures in 1976 and then increasing exponentially to the present count since around 1994. In this short talk, we shall demonstrate how computational researchers use the PDB in proteomics research. We shall show how to search for particular protein structures or groups of structures that possess certain user-defined criteria, visualize protein 3D structures, analyze them, download the corresponding coordinate file(s) – called PDB files – of user-selected
structure(s) and identify the contents of a PDB file by analyzing its format to understand what each of its dozen or so fields (columns) mean and what each of its possibly thousands of records (rows) represent. We shall also discuss the concept of structure-based function prediction (SBFP) of proteins and why it is increasingly becoming the paradigm in proteomics. Finally we shall discuss some applications of SBFP by using as illustrative examples some of the author’s recent works on the following specific topics: (a) the prediction of specific ligand binding sites in proteins and its implementation; (b) the prediction of protein-protein interaction partners and its implementation; (c) the quantification of ligand binding site burial in proteins and its implementation; and (d) the transformation of a PDB file (which is in Cartesian coordinates) into spherical coordinates and its special uses and its implementation. If time permits we shall also demonstrate how to analyze a protein structure by manipulating its PDB file using the Fortran 77/90 language.

3:55-4:15 A novel and quick method to power pilot studies for the comparison of assay platforms under controlled specificity

Zhuolin He, Ziqiang Chen, and Daniel Gaile
Department of Biostatistics, SUNY at Buffalo
E-mail: zhe4@buffalo.edu

Estimating and comparing the “Capability of Detection” has become increasingly relevant with the rapid development and commercialization of novel bioassay platforms utilizing nano- and other small scale technologies. One can follow the International Organization for Standardization (ISO) guidelines for assessing the “Capability of Detection” (ISO-11843) in experimental settings. However, the guidelines do not address the performance under small sample sizes nor the design/analysis of comparative studies. Our work addresses this notable gap. We consider the design of two device/two concentration (2D/2C) pilot experiments, wherein one concentration is assumed to be blank and the other is chosen based upon experimental design considerations, such as power. This problem can also be related to the binormal ROC literature. Comparisons among existing methods are conducted, and results are obtained via numerical integration and Monte Carlo studies. In small sample settings, the “Limit of Blank” (LoB) cut-offs suggested by ISO do not control false positive rates as intended, and naive application of tests conditioning on LoB estimators can be anti-conservative with respect to type I error control. Corrected estimators for LoB are provided via numerical integration. A computationally efficient bootstrap-based test is developed and it is shown to control the type I error probability. Our talk will make special note of two interesting findings: (1) the loss of type I error control does not diminish with increasing sample size as one might expect, and (2) existing methods to compare the sensitivities of two biomarkers for a “fixed” specificity do not always control the specificity (as our approach does).

SESSION 5D

Methods to Enrich Statistics Education

3:30-3:50 Enriching statistics classes in K-12 schools by adding practical ideas from the history of mathematics and statistics

Mucahit Polat and Celal Aydar
Graduate School of Education, SUNY at Buffalo
E-mail: mucahitp@buffalo.edu

As we learn from our experiences we can learn from history as an accumulation of billions of people’s experiences; and as we have joy and excitement from our personal memories, we can have joy and excitement from memories of those people who lived before us. That is why history could also be used as a powerful engagement and meaning-making tool when teaching mathematics and statistics. Throughout history, mathematicians have come up with many creative algorithms and interesting methods to solve problems. Today we are just using and teaching selected and refined methods to solve mathematical problems and to do calculations. Many of the historical methods could be used in the classroom to enrich the learning experience for students. Teaching mathematics and statistics accompanied by history helps students make meaning and connect concepts. Without historical connections, students may perceive mathematics and statistics as a set of discrete topics and this could prevent them from seeing the connectedness and relevance of topics within mathematics. History also gives mathematics a human face by enabling students to see themselves as part of the general cultural and social processes. This presentation will provide multiple practical ideas that could be used as enrichment and/or as part of a statistics lesson in K-12 schools.
Some of these ideas include: How did Athenians use central tendency to calculate the height of a castle? How did Al-Kindi use statistics and frequency to encrypt messages? Who first introduced and studied the normal curve? How did Laplace estimate the population of France in the 18th century? Who first used the term “variance” and for what purpose? How did Florence Nightingale use statistics to address public health care issues? Who is the founder of Bayes’ theorem, Thomas Bayes or Pierre-Simon Laplace? Why did Fisher reject the Bayesian view during the early 1900s?

3:55-4:15  Multiple linear regression with sports data §

Matthew D’Amico, Jake Ryder, and Yusuf Bilgic
Department of Mathematics, SUNY Geneseo
E-mail: md24@geneseo.edu

We will share our designed interactive lesson for teachers teaching multiple linear regression to students in introductory statistics courses. The interactive lesson is built with R tools used to predict sports statistics based on a variety of explanatory variables. We will investigate finding relationships between explanatory variables like the age of an athlete, the statistics compiled both in individual past years and as a career by that athlete, increases or decreases year to year in production (and more), and response variables such as rushing statistics (yards and touchdowns), passing statistics, and receiving statistics (and more). We will focus primarily on offensive players, as statistics are more readily available, but are open to looking into defensive statistics such as tackles, sacks, and interceptions if obvious trends develop. Through use of these exercises, we will identify trends and their significance between explanatory variables and response variables individually and in groups (e.g., age and rushing yards from the season before predicting rushing yards for the next season). This lesson is targeted for an AP/introductory statistics class. This talk will also briefly reference how to create a lesson with R tools.

SESSION 5E

Extreme Values / Sparse Signal Recovery

3:30-3:50  EXTREME statistics: data analysis for disasters §

Nicholas LaVigne
Department of Mathematics, SUNY Geneseo
E-mail: nsl3@geneseo.edu

Probability and statistics are important to more than just scientific claims. When making decisions in domains where risk and uncertainty are involved, special attention must be given to extreme events, good and bad. This corresponds to estimating the tails of probability distributions instead of their centers, e.g., the upper and lower quantiles instead of the mean. Problems occur because most often the relevant quantiles to estimate fall outside the range of data. Consequently, the tools and methods used in conventional data analysis are no longer useful when considering extreme events. Theoretical results such as the generalized central limit theorem, heavy tailed distributions, and extreme value distributions will be discussed along with practical statistical techniques for exploratory data analysis and parameter estimation. Results will be applied to a data set on environmental and industrial disasters in the United States. The general philosophical approach and practical implications will be stressed.

3:55-4:15  Minimax optimal sparse signal recovery with Poisson statistics

Mohamed Rohban
Broad Institute of Harvard and MIT
E-mail: mh.rohban@gmail.com

We are motivated by problems that arise in a number of applications such as online marketing and explosives detection, where the observations are usually modeled using Poisson statistics. We model each observation as a Poisson random variable whose mean is a sparse linear superposition of known patterns. Unlike many conventional problems, observations here are not identically distributed since they are associated with different sensing modalities. We analyze the performance of a Maximum Likelihood (ML) decoder, which for
our Poisson setting involves a non-linear optimization yet is computationally tractable. We derive fundamental sample complexity bounds for sparse recovery when the measurements are contaminated with Poisson noise. In contrast to the least-squares linear regression setting with Gaussian noise, we observe that in addition to sparsity, the scale of the parameters also fundamentally impacts $L_2$ error in the Poisson setting. We show tightness of our upper bounds both theoretically and experimentally. In particular, we derive a minimax matching lower bound on the mean-squared error and show that our constrained ML decoder is minimax optimal for this regime.

**POSTER SESSION**

**Variable bandwidth local polynomial smoothing via local cross-validation §**

Katherine Grzesik and Derick Peterson  
Department of Biostatistics and Computational Biology, University of Rochester  
E-mail: Katherine_Grzesik@urmc.rochester.edu

Nonparametrically estimating a regression function with varying degrees of smoothness or heteroscedasticity can benefit from a smoother that is more flexible than a constant bandwidth local polynomial to efficiently capture its features. We propose estimating a smooth variable bandwidth function using a form of local cross-validation (LCV) based on smoothing the squared leave-one-out cross-validated residuals (SLOORs). Further, we propose a method of blending LCV fits of multiple polynomial orders, again guided by the smoothed SLOORs. The proposed method is computationally quicker than some current methods for functions with sharp changes in smoothness and can yield a reduction in mean squared error for moderate sample sizes while remaining competitive with larger samples.

**Longitudinal associations among three cognitive processes and childrens’ mathematical ability in early elementary school §**

Helyn Kim, Chelsea Duran, Claire Cameron, and David Grissmer  
Curry School of Education, University of Virginia  
E-mail: helynk@gmail.com

In recent years, three cognitive processes – fine motor skills, attention, and visuo-motor integration – have been linked to early and long-term mathematics achievement. These processes, each with motor underpinnings, have been featured centrally in studies assessing the possible role of motor skills in mathematics learning. Yet, beyond well-established associations among these processes, there is little clarity regarding when and to what degree they contribute to mathematical ability in early elementary school. Therefore, this study employed a dynamic systems framework and examined the longitudinal and reciprocal contributions of fine motor coordination, attention, and visuo-motor integration to mathematical ability in a diverse sample of 135 kindergarteners (mean age = 5.6 years, SD = 0.37) and 119 first graders (mean age = 6.7 years, SD = 0.43) over the course of two school years, using an auto-regressive, cross-lag (ACL) approach. The three cognitive processes were measured using the NEuroPSYchological assessment battery, and childrens’ mathematical ability was assessed using the Key Math 3 Diagnostic Assessment. Results from the ACL models revealed longitudinal associations between the three cognitive processes were transactional in nature, with more reciprocal transactions occurring in kindergarten than in first and second grades. These three processes also appeared to be intricately linked to mathematical ability, with cross-lagged effects strongest in kindergarten and diminishing over time. Our findings have important implications for research and practice. Breaking down a more complex skill to identify the components and skill trajectories that underlie mathematics, as well as how these skills are interrelated and build upon each other, is important, not only for understanding the underlying mechanism, but also for informing how best to foster their development in young children.

**Gamma band auditory steady-state response phase-locking angle abnormality in schizophrenia §**

Brian Roach  
Department of Statistics, Texas A&M University  
Judith Ford and Daniel Mathalon  
Department of Psychiatry, University of California at San Francisco  
E-mail: brian.roach@email.tamu.edu

In 1999, Kwon et al reported several gamma band abnormalities in auditory steady state responses (ASSRs) in schizophrenia prompting ~100 subsequent studies across species. While reductions in ASSR power and phase-locking factor (PLF) have been replicated, phase delay was never tested again. We compared differential sensitivity of power, PLF, and phase locking angle (PLA) in
schizophrenia. EEG data were recorded from 28 schizophrenia patients (SZ) and 25 healthy controls (HC) listening to click-trains to obtain ASSRs. Total power, PLF and PLA were calculated after Morlet wavelet time-frequency decomposition of single trial data from electrode Fz. All three time-frequency measures were standardized based on the HC group data to facilitate comparisons between measurements. Power, PLF, and PLA measures were reduced in SZ, with PLA reductions being greatest and unrelated to reductions in power and PLF. PLA correctly differentiated SZ from HC with 85% accuracy compared to 67% and 65% for power and PLF, respectively. PLA is more sensitive to schizophrenia than time-frequency measures typically used to assess 40 Hz ASSR abnormalities. PLA provides a unique neurobiological measure of gamma band abnormalities that characterize SZ and may reflect a different underlying pathophysiological mechanism than PLF and power.

Nonnegative matrix factorization with Zellner penalty

Matthew Corsetti
School of Mathematical Sciences, Rochester Institute of Technology
E-mail: corsetti.matthew@gmail.com

Nonnegative matrix factorization (NMF) is a relatively new unsupervised learning algorithm that decomposes a nonnegative data matrix into a parts-based, lower dimensional, linear representation of the data. NMF has applications in image processing, text mining, recommendation systems and a variety of other fields. Since its inception, the NMF algorithm has been modified and explored by numerous authors. One such modification involves the addition of auxiliary constraints to the objective function of the factorization. The purpose of these auxiliary constraints is to impose task-specific penalties or restrictions on the objective function. Though many auxiliary constraints have been studied, none have made use of data-dependent penalties. In this paper, we propose Zellner nonnegative matrix factorization (ZNMF), which uses data-dependent auxiliary constraints. We assess the facial recognition performance of the ZNMF algorithm and several other well-known constrained NMF algorithms using the Cambridge ORL database.

Investigating toxicology results among suicide victims in upstate New York

Kristen Berish and Jennifer Bready
Division of Math and Information Technology, Mount Saint Mary College
E-mail: kber3040@my.msmc.edu

Suicide is the tenth leading cause of death in the United States. National data, collected by the CDC, and other research show many trends among suicide victims. In our current study the data, collected from Dutchess County over 12 consecutive years, include 312 suicide victims with 241 males and 71 females ranging in age from 14 to 94. The data were broken down into categories defined by age, race, method of suicide, and toxicology results. The primary aim of this study was to find and discuss the links between suicide and alcohol consumption and drug use, both individually and collectively, by examining the victim’s postmortem toxicology results. Postmortem toxicology tests of suicide victims are used to identify the presence of toxins, which include alcohol, prescription drugs, and drugs of dependence that may have been taken prior to their death. In recent years, the connection between suicides and alcohol consumption and drug use have been commonly examined through research. The results of our research discovered that, consistent with previous studies, 33% of the victims had alcohol in their system before death, 14% of the victims had illicit drugs present, and 56% had prescription drugs present in their system. As a whole, females were more likely than males to test positive for toxins in their system. Compared to victims of death by gunshot or suffocation/hanging, victims of poisoning and other methods were more likely to test positive for toxins in their systems. However, we would expect that all victims of death by poison would have drugs in their system. Further research will be conducted to analyze data on the types of drugs used and determine if levels of prescription drugs were within normal limits.

Distance estimation in the presence of quasi-latent structure, with application to methylation array data

Sarah Kaus, Wenjun He, and Daniel Gaile
Department of Biostatistics, SUNY at Buffalo
Devin Koestler
Department of Biostatistics, University of Kansas Medical Center
Xuefeng Ren
Department of Epidemiology and Environmental Health, SUNY at Buffalo
E-mail: sarahkau@buffalo.edu

Numerous approaches for the analysis of DNA methylation array data utilize feature to feature (e.g., probe to probe) and/or sample to sample distance estimators either explicitly (e.g., formal clustering and dimension reduction techniques) or implicitly (e.g., smoothing
methods and heuristics to identify significant contiguous genomic regions). Despite the prevalence of such approaches in the DNA methylation literature, the extent to which such methods are affected by quasi-latent structures in methylation data remains unclear. We explore the plausibility of quasi-latent structure existing in methylation data with the objective of determining the extent to which it impacts statistical inference. We begin by considering a simple mixture model case for bivariate normal distributions and examine how well common distance metrics applied to the manifest variables (i.e., beta values) agree with distance metrics applied to the latent state variables. We extend our work to DNA methylation array data by fitting beta mixture models to two methylation array datasets (i.e., GSE50660 and a select subset of the Breast Cancer cohort from The Cancer Genome Atlas). We demonstrate that, under simple bivariate normal mixture models, correlation based dissimilarity metrics can be misleading. Under some conditions, Euclidean and Manhattan distance estimates can provide reasonable approximations of the underlying distance between latent state distributions. We provide examples of real gene level methylation profiles that display that our findings are relevant in the context of methylation array data. We demonstrate that naive application of methods that assume that a quasi-latent structure is not present can lead to losses in power/efficiency/performance.

**High dimensional variable selection in a prostate cancer study** §

**Chi Chen and Jiwei Zhao**  
Department of Biostatistics, SUNY at Buffalo  
E-mail: chichen@buffalo.edu

This project focuses on the selection of variables (gene types) from high dimensional data in a prostate cancer study by Scott A Tomlins (2007). The GEO dataset GDS3289 from NCBI including 104 samples (specific cell populations) and 20,000 predictor variables (gene types), as well as disease status of each sample, is the raw data in the project. The disease status is treated as the response and gene types are treated as predictors. Although there are no missing values among 1,894 columns of all the 20,000 gene types, it is still risky to perform a logistic regression directly with these variables due to the large number of predictors and small sample size. Our strategy is to minimize the negative binomial log-likelihood, which is also subject to the restriction that the sum of absolute coefficients is less than a given constant. Specifically, penalized logistic regression is formulated in order to select significant nonzero coefficients and corresponding variables. A cross-validation method is applied to search for the estimate of the best restriction limit of the sum of absolute coefficients. There are 5 different measures involved in the LASSO method: actual deviance, squared loss, mean absolute error, misclassification error, and area under the ROC curve. A robust algorithm based on frequency is also applied in order to reduce the randomness of the LASSO result. In addition, minimax concave penalty (MCP) and smoothly clipped absolute deviation (SCAD) penalties are adopted for comparison with the LASSO. Finally, all of the significant nonzero coefficients estimated by means of 7 different methods are summarized, where the LASSO minimizing the mean absolute error selected the largest number of variables (41) while the MCP selected the least number of variables (6).

**Modeling impacts of global financial crises on stock volatility of Nigerian banks** §

**Maruf Raheem and Patrick Ezepue**  
Department of Engineering and Mathematics, Sheffield Hallam University  
E-mail: rahemarsac@yahoo.com

This research aimed at determining most appropriate heteroscedastic model for predicting volatility of 10 major Nigerian banks: Access, United Bank for Africa (UBA), Guaranty Trust, Skye, Diamond, Fidelity, Sterling, Union, ETI and Zenith banks using daily closing stock prices of each of the banks from 2004 to 2014. The models employed include ARCH (1), GARCH (1, 1), EGARCH (1, 1) and TARCH (1, 1). The results show that all of the banks’ returns are highly leptokurtic, significantly skewed and thus non-normal across the four periods except for Fidelity bank during financial crises, findings similar to those of other global markets. There is also strong evidence for the presence of heteroscedasticity, and that volatility persistence during crisis is higher than before the crisis across the 10 banks, with that of UBA taking the lead, about 11 times higher during the crisis. Findings further revealed that asymmetric GARCH models became dominant especially during financial crises and post crises when the second reforms were introduced into the banking industry by the Central Bank of Nigeria (CBN). Generally one could say that Nigerian banks’ returns are volatility persistent during and after the crises, and characterized by leverage effects of negative and positive shocks during these periods.

**A multivariate logistic regression analysis for spousal concordance in health behavior change** §

**Zehua Zhou and Jiwei Zhao**  
Department of Biostatistics, SUNY at Buffalo  
E-mail: zhuazho@buffalo.edu

This project focuses on the selection of variables (gene types) from high dimensional data in a prostate cancer study by Scott A Tomlins (2007). The GEO dataset GDS3289 from NCBI including 104 samples (specific cell populations) and 20,000 predictor variables (gene types), as well as disease status of each sample, is the raw data in the project. The disease status is treated as the response and gene types are treated as predictors. Although there are no missing values among 1,894 columns of all the 20,000 gene types, it is still risky to perform a logistic regression directly with these variables due to the large number of predictors and small sample size. Our strategy is to minimize the negative binomial log-likelihood, which is also subject to the restriction that the sum of absolute coefficients is less than a given constant. Specifically, penalized logistic regression is formulated in order to select significant nonzero coefficients and corresponding variables. A cross-validation method is applied to search for the estimate of the best restriction limit of the sum of absolute coefficients. There are 5 different measures involved in the LASSO method: actual deviance, squared loss, mean absolute error, misclassification error, and area under the ROC curve. A robust algorithm based on frequency is also applied in order to reduce the randomness of the LASSO result. In addition, minimax concave penalty (MCP) and smoothly clipped absolute deviation (SCAD) penalties are adopted for comparison with the LASSO. Finally, all of the significant nonzero coefficients estimated by means of 7 different methods are summarized, where the LASSO minimizing the mean absolute error selected the largest number of variables (41) while the MCP selected the least number of variables (6).

This research aimed at determining most appropriate heteroscedastic model for predicting volatility of 10 major Nigerian banks: Access, United Bank for Africa (UBA), Guaranty Trust, Skye, Diamond, Fidelity, Sterling, Union, ETI and Zenith banks using daily closing stock prices of each of the banks from 2004 to 2014. The models employed include ARCH (1), GARCH (1, 1), EGARCH (1, 1) and TARCH (1, 1). The results show that all of the banks’ returns are highly leptokurtic, significantly skewed and thus non-normal across the four periods except for Fidelity bank during financial crises, findings similar to those of other global markets. There is also strong evidence for the presence of heteroscedasticity, and that volatility persistence during crisis is higher than before the crisis across the 10 banks, with that of UBA taking the lead, about 11 times higher during the crisis. Findings further revealed that asymmetric GARCH models became dominant especially during financial crises and post crises when the second reforms were introduced into the banking industry by the Central Bank of Nigeria (CBN). Generally one could say that Nigerian banks’ returns are volatility persistent during and after the crises, and characterized by leverage effects of negative and positive shocks during these periods.
Behaviors of using preventive health services such as cholesterol screening and getting a flu shot can have a positive influence on people’s health. One critical question is how to encourage people to make decisions regarding health utilization to improve public health. Some literature indicates that among married couples, spousal behavior is considered as a critical factor in adopting, continuing, or relapsing to poor health behaviors. Thus, we’d like to examine the effect of spousal behavior change on shaping individuals’ decisions to use preventive care as well as the association between them. Our data are from the Health and Retirement Study (HRS, 1992), a longitudinal study with a nationally representative sample of individuals (born between 1931 and 1941) and their spouses (regardless of age). By building a series of logistic regression models, we explore the association between the respondents’ and their spouses’ behavior changes. Our main covariate is the spousal behavior change, containing 4 possible values: starting, continuing, stopping and never participating. Since we restrict the sample to those who have “less healthy” behavior at the beginning, our dependent variable is whether an individual initiates behavior change. Moving from an unadjusted model to the fully adjusted model, we incrementally add sets of variables including demographic control variables, life event variables, health status variables and health change variables. We consistently find that for cholesterol screening, having a spouse who starts screening is positively associated with the individual also starting screening, compared with having a spouse who never initiates screening. Similarly, having a spouse who initiates getting a flu shot strongly encourages the individual to start getting vaccinated compared with having a spouse who never gets a flu shot. These results potentially help to develop interventions, treatments and policies to improve health habits, thereby improving public health.

Evaluation of nonparametric distance-based testing methods for the analysis of multivariate data with partial responses §

Jingchen Zhang, Rachael Blair, and Daniel Gaile
Department of Biostatistics, SUNY at Buffalo
E-mail: jzhang67@buffalo.edu

We consider four multivariate distance based tests for the comparison of two samples: (1) the pseudo-$F$ statistic from a nonparametric multivariate analysis of variance, (2) Biswas and Ghosh’s interpoint distance test, (3) the truncated and non-truncated multi-response permutation procedures test (MRPP) of Mielke, and (4) the Cramer test. While the performance of these methods has been evaluated with respect to location and scale alternatives, we consider partial response alternatives. We compare the performance of the tests under a variety of bivariate and multivariate normal simulation conditions. We apply the multivariate tests to Euclidean as well as rank-based distance matrices. We also compare the performance of these tests to Bonferroni-corrected univariate tests as well as Hotelling’s $T^2$ test. We apply the methods to a gene expression data set from which gene signatures related to Tuberculosis were identified and validated across three clinical subpopulations. We provide R functions for the rapid application of the considered methods. We demonstrate that the Cramer test is identical to a version of the pseudo-$F$ statistic. We provide simulation conditions for which the distance-based tests outperform the univariate and Hotelling’s $T^2$ tests. We demonstrate that Mielke’s truncated version of the MRPP does not deliver advantages in the presence of partial response, nor does a simple trimming procedure. Under the simulation settings we considered, the interpoint test of Biswas and Ghosh was never superior to, and was sometimes outperformed by, other tests, the pseudo-$F$ statistic most notably. We demonstrate that a distance-based testing approach can offer advantages over a cluster enrichment approach.

K-levels leave one out filtering: a novel feature selection technique with application to methylation array data §

Wenjun He and Daniel Gaile
Department of Biostatistics, SUNY at Buffalo
Devin Koestler
Department of Biostatistics, University of Kansas Medical Center
Xuefeng Ren
Department of Epidemiology and Environmental Health, SUNY at Buffalo
E-mail: wenjunhe@buffalo.edu

Epigenome-wide association studies (EWAS) using DNA methylation microarrays frequently suffer from what has been termed “the curse of dimensionality”. One strategy for the analysis of such data is to identify and analyze a “feature selected” subset of probes. Common approaches utilize variance, goodness of fit, and model based filters. We propose a novel method. We evaluate the performance of filtering techniques using publically available methylation array data (i.e., GSE50660) from a well-designed and executed study involving former and non-smokers from three separate age groups. We identify probes that are specific for differentially methylated regions (DMRs) with respect to the full datasets. We then explore the effectiveness of filtering techniques to provide feature subsets that are enriched for those DMRs. We consider feature selections applied to the full data as well as random subsets of the data (i.e., randomly selected to mimic less powered studies). We propose and evaluate a novel method, which we refer
to as “K-levels leave one out” filtering (KLOO). Variance-based filtering applied to beta values outperforms those applied to logit transformed values. Variance-based filters also outperform modality based filters based upon the Dip Test and a mixture model approach. Some of the significant DMRs provide evidence of a partial response deviation from the tested null hypothesis and we relate this finding to our filtering results. We demonstrate how heteroscedasticity in the data affects the filtering techniques. We present simulation results for our KLOO filter. Application of KLOO to the smoking/age methylation dataset provides proof of principle that it can be an effective filtering technique. Experimental design considerations are also presented.

A dual frame estimator considering the adjustment for non-response in telephone surveys

Taeyoung Kim
Educational Psychology and Quantitative Methods Program, School of Education, SUNY at Buffalo
E-mail: tkim33@buffalo.edu

As cell phone use in the United States grows, the potential for coverage bias in random digit dial (RDD) telephone surveys will increase if these surveys continue to exclude most cell phone numbers. In this regard, an alternative approach like dual frame sampling needs to be implemented to cover the entire population. The present study first examined several estimators involved in surveys using sampling frames for both landlines and cell phones. Meanwhile, the non-response issue is still outstanding even after adopting these dual frame sampling schemes. The goal of this study is to provide another estimator which is based on post-stratification when non-response occurs, and to compare the performance of the suggested estimator with existing ones. The study briefly covered the characteristics of the post-stratification estimator and its variance estimator, and conducted a series of comparisons via simulations. The dual frame estimators introduced in the present study include the ones suggested by Hartley, Fuller & Burmeister, and the pseudo maximum likelihood (PML) estimator, which is the extension of the ML method under the complex design. The author employed different conditions in terms of response patterns generated by different underlying models, keeping the average response rate identical across different conditions to reflect various situations in practice. The results suggest that relative bias and relative standard error of the post-stratification estimator are better than those of other methods. Finally, this study concludes with applications of the non-response adjusted post-stratification estimator involving a dual frame setting to relevant disciplines.

Exact unconditional test for the comparison of two proportions under negative binomial sampling

Shiva Dibaj, Alan Hutson, and Gregory Wilding
Department of Biostatistics, SUNY at Buffalo
E-mail: seyedehs@buffalo.edu

Comparing two independent binomial samples in a $2 \times 2$ contingency table can be done by asymptotic or exact methods. However, in studies where the outcome of interest is rare, to ensure the final data contain an adequate number of events, samples might be collected by inverse sampling. Statistical inferences proposed in this context have been mostly asymptotic-based procedures. However, large-sample approximations for distributions of parameter estimates and statistics when dealing with discrete data are not the best option especially when sample sizes are not large. An alternative method for analyzing data from such tables is an exact procedure. We develop an unconditional exact test to compare two proportions under negative binomial (NB) sampling and then extend our exact method by Berger and Boos’ modification. The two most common statistics in large sample theory, likelihood ratio and Wald’s test statistics, are used. The exact p-value for each possible table in the sample space is defined as the summation of the probabilities of tables that are no more likely than the observed one. In order to eliminate the NB nuisance parameter we take the supremum over the entire range of the parameter in exact method and over a 99.99% confidence interval for the parameter. In a vaccine efficacy trial, the expected number of samples using our method is compared to the required sample size with conventional methods. Results show that while conventional methods require a much larger number of samples to achieve the desired power level, they have low probability of observing enough events of interest.

Penalized pairwise pseudo-likelihood for variable selection with missing data

Yang Yang and Jiwei Zhao
Department of Biostatistics, SUNY at Buffalo
E-mail: yyang39@buffalo.edu

The importance of variable selection in data analysis speaks for itself with its popularity in the statistical literature. However, there are many obstacles preventing us from correctly selecting variables. Missing data, especially non-ignorable, in the response variable, covariates, or both, represent a major difficulty that we need to overcome. In particular, discussion of missing data in a high-dimensional setting is rarely seen. Conventional approaches are able to build corresponding models on the basis of knowing the correct missing data mechanism, but without a known missing data mechanism, they don’t work very well. To cope with this
difficulty, we propose an alternative variable selection procedure based on pairwise pseudo-likelihood. After converting pseudo-likelihood to a regular likelihood function of logistic regression without the intercept, it becomes straightforward for us to add some common penalties, such as the LASSO, SCAD, and MCP, to achieve the goal of variable selection. In practice, we use the R package “glmnet” to perform variable selection for LASSO, and an iteration method for SCAD and MCP. We adopt cross-validation to choose the tuning parameter. The main advantages of our variable selection procedure are that no parametric assumptions are needed on the missing data mechanism and it works in both low dimension \((n > p)\) and high dimension \((n \leq p)\) situations. Also, simulation results show that, even if there is a weak signal in the original parameter and the percentage of missing data is high (about 35%), our method demonstrates great improvement in variable selection over some existing naïve methods.

**A novel exact method for top-K family-wise error control for small sample non-small scale inference §**

Ziqiang Chen and Daniel Gaile  
Department of Biostatistics, SUNY at Buffalo  
Devin Koestler  
Department of Biostatistics, University of Kansas Medical Center  
Xuefeng Ren  
Department of Epidemiology and Environmental Health, SUNY at Buffalo  
E-mail: ziqiangc@buffalo.edu

We consider a two stage study in which \(M\) feature-level assay values are measured for each of \(N\) samples in the first stage and the top \(K\) features are identified and carried forward to a second stage validation study. We consider the case in which the top-K features are identified via p-value ranking of a discrete univariate test applied to each of the \(M\) features. Our interest is in providing type I error control for a formal test of the hypothesis that at least one of the top-K features is, in fact, non-null. We provide a seminal exact test for the hypothesis that at least one of a list of top-K features is truly significant. We demonstrate that, under certain conditions, our method can provide a more practical alternative (and notably superior in some ways) to conventional FWER, K-FWER, and FDR approaches. We elucidate the conditions under which the application of our method is feasible. Such conditions are largely dictated by issues pertaining to experimental design and computational feasibility. We apply our method to publicly available methylation array datasets and provide proof of principle of its utility in real world settings.

**Estimating the empirical null distribution of maxmean statistics in gene set analysis §**

Xing Ren and Jeffrey Miecznikowski  
Department of Biostatistics, SUNY at Buffalo  
Jianmin Wang and Song Liu  
Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute  
E-mail: xingren@buffalo.edu

Gene Set Analysis (GSA) is a framework for testing the association of a set of genes and the outcome, e.g., disease status or treatment group. The method relies on computing a maxmean statistic and estimating the null distribution of the maxmean statistic via a restandardization procedure. In practice the pre-determined gene sets have stronger intra-correlation than genes across sets. This is not taken into account in the restandardization procedure and can result in biases in the estimated null distribution. We derive an asymptotic null distribution of the maxmean statistic based on large sample theory. We propose a flexible two group mixture model for the maxmean statistic. Under reasonable assumptions, the mixture model allows us to estimate the null parameters empirically via a maximum likelihood approach. Our empirical method is compared with the restandardization procedure of GSA in simulations. We show that our method is more accurate in null density estimation the false discovery rate control when the genes are strongly correlated within gene sets.

**Using R for analysis of basketball results §**

Steven Treacy and Reneta Barneva  
Department of Applied Professional Studies, SUNY Fredonia  
E-mail: trea4547@fredonia.edu

Records about sports have always been kept. For some popular sports large data sets have been created, which experienced unprecedented growth in the Digital Era. With the means of data analytics, information can be extracted from these data sets and can be used for various purposes such as assessing and comparing the players, teams, games, and others. Unfortunately, data analytics is rarely taught to sport management students, prospective coaches or athletes. The available tools and languages are perceived as
requiring specialized knowledge in mathematics, statistics, or computer science and are rarely used by sport-related people. At the same time, professional sports teams and sponsors hire professional data analysts to help them make their decisions. In this work, we will demonstrate how the R language could be used by even non-professionals to explore sports-related data. R is a language for which there exists a free development environment running on the main platforms. It provides various statistical tools, but its main strength that sets it apart from other statistical packages and systems is the visualization techniques: it is able to provide various plots that stimulate the intuition of the sports experts. In this project we illustrate how R could be used even by non-professionals for sports analysis. We take a sport, Basketball for example, and make a hypothesis, such as that the newer basketball players are performing better than the older basketball players. Using information such as total points scored, free throws, two-point shots, and three-point shots, we could compile all that data on certain individuals, such as the Most Valuable Players (MVPs), and compare them across different teams. The analysis will show whether those MVPs indeed either score higher or make harder shots more consistently.

**Obesity and its correlation with non-natural deaths in Dutchess County, NY**

Joseph Ianniello  
Division of Natural Sciences, Mount Saint Mary College  
E-mail: jian5434@my.msmc.edu

According to the National Institutes of Health, the increasing rates of individuals being obese and overweight, together, have accounted for the second leading cause of preventable deaths in the United States, close behind tobacco use. An estimated 300,000 deaths per year are due to the obesity epidemic. Approximately 30% of adults residing in New York State are classified as being either overweight or obese. The National Heart, Lung, and Blood Institute provided the body mass index (BMI) categories used to classify individuals for this study as: underweight, normal weight, overweight, or obese. Health effects of being obese or overweight have been heavily researched, with resulting health issues stemming from: cardiovascular disease, diabetes mellitus, hypertension, etc. However, this study aimed to examine rates of non-natural deaths in obese and overweight individuals by analyzing over 600 cases of death reported by the Dutchess County Medical Examiner’s Office. Variables analyzed include: gender, age, height, weight, BMI, heart and liver weights, panniculus layer, and manner and cause of death. The presenter will discuss the statistical analysis of the non-natural death data, make comparisons among all four BMI categories and their association to these non-natural deaths, and will share why this statistical analysis is important to the county (and other counties).

**Knowledge guided data driven method for patient profiling**

Yuqing Xing  
Department of Statistics, University of Massachusetts – Amherst  
E-mail: yuqingxing88@gmail.com

The traditional standardized care relied on a “one-size-fits-all” approach that proved to be not very effective in the treatment and intervention for many individual patients. To improve effectiveness, we need to more precisely profile the characteristics of the individual patient or small population. Since healthcare data are typically noisy, sparse, high dimensional, and highly correlated, the raw features typically are not the best way to profile a patient as they lack insight for action. It is desired to reduce a large number of raw features to a few meaningful latent variables for better interpretation and improved modeling accuracy. In addition, leaving the abundant medical knowledge behind, the purely data-driven approaches sometimes lead to rediscovery of what is already known. In our work, we are leveraging existing medical knowledge such as ontology to organize raw features into many medically meaningful groups, e.g., lifestyle-related, mental health-related, and diabetes-related groups. Next we train different Restricted Boltzmann Machines (RBMs) to represent each medically meaningful group. We then combine hidden representations generated from different RBMs as a single input for the second RBM layer. We can stack multiple RBMs in this way to form a deep structure. The final representation from this structure will be used to profile individual patients in different applications.

**Big data in the music industry**

Tylce Lewis  
Department of Computer and Information Sciences, SUNY Fredonia  
Reneta Barneva  
Department of Applied Professional Studies, SUNY Fredonia  
E-mail: lewi5087@fredonia.edu

The development of computing technology has changed drastically the ways music is listened to, musical events are organized, and marketing and ticket sales are performed. Until recently, the Internet was viewed as an adversary of the music industry and there were many disputes over copyright and illegal downloads. In recent years, however, this trend has changed and now companies such as
Spotify and YouTube that offer free streaming music are thriving. The reason for this is that they were able to collect data, successfully apply data analytics, and find a way to use the extracted information. In the past, the music companies had a vague idea about the buyers of their discs. With music streaming, large databases were created tracking the preferences of the users, the time and periods they are listening, and the number of downloads of a musical piece. For example, the Musical Genome Project has analyzed over 30 million songs and has classified them according to 450 characteristics such as the speed of the rhythm, the type of instruments, the gender of the vocalists, and others. The information is used by Pandora to track user preferences and make recommendations. Pandora's rival, Spotify, already possesses over 28 petabytes of music industry data. Apart from tracking the user's behavior, it collects information about the artists and employs “likes.” This allows it to discover rising stars and predict the positions of the songs in the Billboard charts. Our presentation is an overview of the various ways Big Data Analytics can be used in music industry and of the statistical methods used.

**Modeling food policy decision analysis with Bayesian networks**

Rachel Wilkerson  
Texas Hunger Initiative, Baylor University  
E-mail: Rachel_Wilkerson@baylor.edu

The efficacy of policy interventions for socioeconomic challenges like food insecurity are difficult to measure due to a limited understanding of the complex web of causes and consequences. As an additional complication, limited data are available for accurate modeling. Thorough risk-based decision making requires appropriate statistical inference and a combination of data sources. The federal summer meals program is a part of the safety net for food insecure families in the U.S., though the operations of the program itself are subject to risk. These uncertainties stem from variables both about internal operations as well as external food environment. Local partners often incur risk in operating the program; thus we use decision analysis to minimize the risks. After integrating public, private, and government data sources to create an innovative repository focused on the operations of the child nutrition programs, we construct a Bayesian network of variables that determine a successful program and compute the expected utility. Through an expected utility analysis, we can identify the key factors in minimizing the risk of program operations. This allows us to optimize the possible policy interventions, offering community advocates a data-driven approach to prioritizing possible programmatic changes. This work represents substantial progress towards innovative use of government data as well as a novel application of Bayesian networks to public policy. The mathematical modeling is also supplemented by a community-facing application developed in Shiny that aims to educate local partners about evidence-based decision making for the program operations.

§ Indicates presentations that are eligible for the student presentation awards