2019 Lloyd Roeling Conference
Titles and Abstracts
(final - updated 24 October 2019)

Estimation of Hospital Quality Based on Performance Measure and Hospital Acquired Infection Data: A Bayesian Bivariate Approach
Dulal K. Bhaumik
The University of Illinois at Chicago

Hospital quality data consists of dichotomous process measurements and count outcome measures. However, current statistical approaches to evaluating hospital quality often utilize one of these measurements. Moreover, when data are collected from multiple units within a hospital, the within-hospital correlation is typically ignored. To address this, we introduce a novel methodology for determining hospital quality that incorporates both outcomes within a Bayesian framework. We illustrate our approach using performance measures obtained from the Joint Commission and hospital acquired infection measures directly from hospitals. Using our approach, we differentiate high and low performing hospitals with a high degree of certainty.
This work is joint with David J. Morton and Rawan A. Rupnow.

A Measure of How Supportive the Data Is
Seongbaek Yi
Department of Statistics, Pukyong National University, Busan, Korea

There have been many reports from various fields of science on problems and misunderstandings of p-value and statistical significance. As a consequence the American Statistical Association (ASA) published a statement about p-values including online discussions in 2016 and finally devoted entirely to that topic in the March 2019 issue of The American Statistician. Its main message is that the threshold p-value of 0.05 is arbitrary and the notion of being statistically significant" based on the value does not make sense in many situations. Among many alternatives are some suggestions that p-values need to be complemented with confidence intervals. In this paper we propose a complementary index measure of how supportive the data is for hypothesis, based on the length of confidence interval and the location of the specified parameter in the interval. The measure plays the role of p-value and confidence interval as well and furthermore tells the extent to which observed evidence is for the specified model. We consider the measure under various contexts including normal model with known variance and examine its relationship with test statistic, confidence interval and sample size.
This is joint work with Taewoong Uhm.

Fourier transform and project methods in kernel entropy estimation for linear processes
Hailin Sang
Department of Mathematics, The University of Mississippi

Entropy is widely applied in the fields of information theory, statistical classification, pattern recognition and so on since it is a measure of uncertainty in a probability distribution. The quadratic functional plays an important role in the study of quadratic Rènyi entropy and the Shannon entropy.

It is a challenging problem to study the estimation of the quadratic functional and the corresponding entropies for dependent case. In this talk, we consider the estimation of the quadratic functional for linear processes. With a Fourier transform on the kernel function and the projection method, it is shown that, the kernel estimator has similar asymptotical properties as the i.i.d. case studied in Giné and Nickl (2008) if the linear process $\{X_n: n\in \N\}$ has the defined short range dependence. We also provide an application to L^2_2 divergence and the extension to multivariate linear processes. The simulation study for linear processes with Gaussian and α-stable innovations confirms the theoretical results. As an illustration, we estimate the L^2_2 divergences among the density functions of average annual river flows for four rivers and obtain promising results.
This is a joint work with Yongli Sang and Fangjun Xu.

An Asymptotic Conditional Test of Independence in Bernoulli Sequences Using the Number of Runs Given the Number of Successes

Sungsu Kim
Department of Mathematics, University of Louisiana at Lafayette

In this presentation, we present the asymptotic normality of the conditional distribution of the number of runs given the number of successes for a sequence of independent Bernoulli random variables. In our proof, the Frobenius-Harper technique is used to represent the number of runs as the sum of independent and not necessarily identically distributed Bernoulli random variables. Then, an asymptotic conditional test for independence is provided. Our simulation results exhibit that the test based on conditional distribution performs better that one based on unconditional distribution, over the entire range of success probability and first order correlation. In addition, the UMVUEs of the factorial moments and the probabilities of the number of runs are presented in this paper.
This is joint work with C.J. Park.

Multivariate Calibration with Robust Signal Regression

Bin Li
Department of Experimental Statistics, Louisiana State University

Motivated by a multivariate calibration problem from a soil characterization study, we proposed tractable and robust variants of penalized signal regression (PSR) using a class of nonconvex Huber-like criteria as the loss function. Standard methods may fail to produce a reliable estimator especially when there are heavy tailed errors. We present a computationally efficient algorithm to solve this nonconvex problem. Simulation and empirical examples are extremely promising and show the proposed algorithm substantially improves the PSR performance under heavy-tailed errors.

A Ratio Technique for Showing Convergence of Known Distributions to Normality or Non-normality

Subhash Bagui
Department of Mathematics and Statistics, The University of West Florida

This talk presents an elementary technique for deriving the convergence of known discrete/continuous type distributions to limiting normal or non-normal distributions. The technique utilizes the ratio of the pmf/pdf at hand at two consecutive/nearby points. This ratio method is illustrated via a few well-known discrete and continuous distributions. The presentation should be of interest to teachers and students in probability and statistics courses.

The Nonparametric Behrens-Fisher Problem with Dependent Replicates

Akash Roy
Department of Mathematical Sciences, University of Texas Dallas

Statistical comparison of two independent groups are one of the most frequently occurring inference problems in scientific research. Most of the existing methods available in the literature are not applicable when measurements are taken with dependent replicates, for example when visual acuity or any blood parameters of mice sharing the same cage are measured. In all these scenarios the replicates should neither be assumed to be independent nor be observations coming from different subjects. Furthermore, using a summary measure of the replicates as a single observation would decrease precision of the effect estimates and thus decrease the powers of the test procedures. Thus, there is a need for purely nonparametric flexible methods that can be used for analyzing such data in a unified way. Ranking procedures are known to be a robust and powerful statistical analysis tool for which parametric distributional assumptions are doubtful.

So, a solution is proposed for these two sample problems with correlated replicates. The results achieved in our work generalize the ideas on previous attempts for testing the rather strict hypothesis H0: F1 = F2 or even for testing H0: p = 1/2. In comparison to the existing pioneering works, differently weighted estimators of the treatment effect p as well as unbiased variance estimators will be proposed in the current work. Therefore, it is of major interest to estimate the treatment effect and to test whether there is any significant difference between these two groups along with the computation of a confidence interval. Weighted as well as unweighted versions of the estimators of the treatment effects are investigated and their asymptotic distributions are derived in a closed form. Furthermore, major attention will be given to the accuracy of the tests in terms of controlling the nominal type-I error level as well as their powers when sample size are rather small. Here, it will be shown that the distributions of the tests can be approximated using t-distributions with approximated Satterthwaite-Welch degrees of freedom. The degrees of freedom are estimated in such a way that the new methods coincide with the Brunner-Munzel test when single measurements are observed. Extensive simulation studies show favorable performance of the new methods. Application of this method is extensively shown in two different toxicological studies involving small sample sizes and different numbers of dependent replicates per unit.

Study on Sensitivity Analysis of Dropped Objects Hitting on the Pipeline

Hanqi Yu
Department of Mathematics, University of New Orleans

Nowadays, submarine pipelines play an important role in marine oilfield industry. As oil industry gradually moves towards deep sea fields with water depth more than 1000 meters, they are subjected to several threats which can cause failure of the line, such as external impact, mechanical defects, corrosion and natural hazards, from which accidentally-dropped objects have become the leading external risk factor for subsea developments. In this paper, a sample field layout introduced in Det Norske Veritas (DNV) guide rules is selected as the study case with water depth at 100m. There are three different groups of “Flat/long shaped” used in this model, whose weights are less than 2 tones (Case 1), between 2 tones and 8 tones (Case 2) and greater than 8 tones (Case 3), respectively. DNV’s simplified method as well as an in-house tool “DROBS” is used to calculate the hit probability and the difference between results are discussed. Meanwhile, the sensitivity analysis on mass, collision area, added mass coefficient, and drag coefficient of the objects are considered at four damage levels calculated based on limit state function. It is concluded that mass and added mass coefficient have the lowest sensitivity level with increment on probability of damage, the collision area and drag coefficient have the highest sensitivity with decrement on damage probability.

Inference on the drug interaction index based on binary response data

Thomas Mathew
Department of Mathematics and Statistics, University of Maryland Baltimore County

The interaction index is a parameter that is used to quantify the interaction when two or more drugs are combined in a combination therapy. The index is defined based on the deviation of the response of the drug combination, from the expected response obtained under a reference model of no interaction. The well-known Loewe additivity model is often used to derive the interaction index; when the drugs do not interact, the interaction index takes the value one. The drugs act synergistically when the index is less than one, and antagonistically when the index is more than one. In the talk, I will consider binary response data and use the fiducial approach to derive a confidence interval for the interaction index. Accuracy of the proposed solution will be assessed based on estimated coverage probabilities, and will be compared with available large sample solutions based on the expected width of the confidence interval. The results will be illustrated using data from a study to assess interaction in a combination therapy involving the analgesic drugs Tramadol and Acetaminophen.
This is joint work with Xiaoshu Feng and Kofi Adragni.

Prediction Intervals for Hypergeometric Distribution

Shanshan Lv
Department of Statistics, Truman State University

The problem of constructing prediction intervals (PIs) for a future sample from a hypergeometric distribution is addressed. Simple closed-form approximate PIs based the Wald approach, the joint sampling approach, and a fiducial approach are proposed and compared in terms of coverage probability and precision. Construction of the proposed PIs are illustrated using an example.

Accelerate Pseudo-Proximal Map Algorithm

Dao Nguyen
Department of Mathematics, The University of Mississippi

In big data era, we are often faced with the challenges of high-dimensional data and complex models, for which the likelihood is either intractable or very expensive to compute. As a result, the simulation-based inference has drawn much attention since it seems to be the only current solution to many real-world problems. Iterated filtering [10, 9] enables simulation-based inference via model perturbations and gradient approximation through sequential Monte Carlo filtering. Using iterated filtering as an approximation of the forward step of the proximal gradient, Guo [7] maximizes the likelihood function by iterating the pseudo-proximal map. In this paper, we improve on this novel idea by accelerating the process with an additional momentum term. We show that under suitable perturbation policy, the proposed framework converges with an optimal rate for both convex and non-convex likelihood function. We demonstrate the efficiency of the algorithm based on a toy model and a challenging model of a biological network, showing substantial improvement over standard approaches.

Kernel smoothing density estimation when group membership is subject to missing

Wan Tang
Department of Biostatistics and Data Science, Tulane University School of Public Health and Tropical Medicine

The density function is a fundamental concept in data analysis. When a population consists of heterogeneous subjects, it is often of great interest to estimate the density functions of the subpopulations. Nonparametric methods such as kernel smoothing estimates may be applied to each subpopulation to estimate the density functions if there are no missing values. In situations where the membership for a subpopulation is missing, kernel smoothing estimates using only subjects with membership available are valid only under missing complete at random (MCAR). In my talk, I will present several new kernel smoothing methods for density function estimates by combining models of the missing mechanism and/or prediction models of the membership under the missing at random (MAR) assumption. The asymptotic properties of the new estimates are developed, and simulation studies and a real study in mental health are used to illustrate the performance of the new estimates.

Dynamics of Hypoxic Zone in Northern Gulf of Mexico

Yi Zhen
Department of Mathematics, University of New Orleans

Hypoxia is the chemical reaction that occurs when the concentration of oxygen is low. Hypoxia has direct impact on environment of aquatic life and coastal region. Nitrogen is one of essential nutrients for plant growth, but overabundance of nitrogen causes eutrophication and can result in algal blooms. The excessive algae growth would result in depletion of oxygen in water and lead to hypoxia. In this project, it aims to quantify the impact of hydrology and variations in nutrient levels on the development of hypoxia in the Northern Gulf Coast of Louisiana. Data of salinity, temperature varying at different depths and nitrogen level of the Gulf Coast in Louisiana from 1985 to present will be collected. The data will be analyzed by various statistical methods and mathematical models will be developed by multiple regression, ARIMA, and tree-based methods, etc. Variation patterns of nitrogen level and formation of hypoxic areas in the Gulf Coast of Louisiana under extreme weather condition will be investigated and the impact will be evaluated.

Local Limit Theorem for Linear Random Fields

Timothy Fortune
Department of Mathematics, University of Mississippi

In this paper, we investigate the conditions under which a local limit result holds for the linear random field X_j = \sum_{i\in\Z^d}a_i \varepsilon_{j-i}, defined on $\Z^d$ with innovations ε_{i} that are independent and identically distributed with mean zero and finite variance. Let $\Gamma_n^d=[-n,n]^d\cap\Z^d$ for each whole number n. Define the sum $S_n = \sum_{i\in\Gamma_n^d} X_i$, and let $s_n = Var(S_n)$. Building on Shore's previous work, we are able to show that the local limit theorem holds for both short and long memory linear random fields in the sense that the sequence of measures $\sqrt{2\pi} s_n P(S_n \in (a,b))$ of the interval (a,b) converges to Lebesgue measure. That is, \lim_{n\to\infty} \sqrt{2\pi} s_n P(S_n \in (a,b)) = b-a, assuming minimal and reasonable requirements of the innovations.
This is joint work with Megda Peligrad and Hailin Sang.

Rethinking Control Chart Design and Evaluation

William H. Woodall
Virginia Tech

Some practical issues are addressed involving the control of the number of false alarms in process monitoring. This topic is of growing importance as the number of variables being monitored and the frequency of measurement increase. An alternative formulation for evaluating and comparing the performance of control charts is given based on defining in-control, indifference and out-of-control regions of the parameter space. Methods are designed so that only changes of practical importance are to be detected quickly. This generalization of the existing framework makes control charting much more useful in practice, especially when many variables are being monitored. It also justifies to a greater extent the use of cumulative sum (CUSUM) methods.
Reference: Woodall, W. H. and Faltin, F. W. (2019). “Rethinking Control Chart Design and Evaluation,” Quality Engineering 31(4), 596-605.

Normalization in Metagenomic Data

Zhide Fang
Biostatistics Program, Louisiana State University Health Science Center

In this talk, we use a metagenomic dataset to demonstrate the ineffectiveness of certain normalization approaches. Then we describe the steps of conducting simulation according to the characteristics learned from the dataset. The impact of normalization on the differential abundance analysis is also discussed.

Statistical Issues on Analysis of Censored Data Due to Detection Limit
Hua He
Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine

Measures of substance concentration in urine, serum or other biological matrices that fall below the assay limit of detection are pretty common in epidemiological and medical research. When concentrations are under the detection limit (DL), accurate measures cannot be obtained and their values are left censored. Common practice for addressing the censoring issue is to delete or `fill-in' the censored observations in data analysis, which often result in biased or non-efficient estimates. In this talk, I will present some statistical issues on analysis of such censored data when the data are treated either as dependent variable or independent variable. Results from simulation studies, NHANES Study and Bogalusa Heart Study, to investigate the statistical issues and compare different methods will be provided.

Stata and Python Integration
Zhao Xu
StataCorp LLC

Stata is a complete, integrated statistical and data science software package that includes intensive usage in data manipulation, visualization, statistics, and reproducible reporting. It also provides comprehensive programming interfaces for users to extend Stata's features using other programming languages such as C/C++ and Java. In Stata 16, Stata provides tight integration with Python which allows users to embed and execute Python code from within Stata and output Python results within Stata. Additionally, the Stata Function Interface (sfi) Python module allows users to pass data and results between Stata and Python. In this talk, I will introduce Stata's new Python integration features and will use them to demonstrate how to interact Stata and Python.

Estimation Similarity Measures and their Sampling Distributions
Madhuri S. Mulekar
Department of Mathematics & Statistics, University of South Alabama

The extent of similarity between two populations is measured using indices of similarity or similarity measures. They are often used to make comparative inferences about two groups, such as describing the degree of inter-specific encounter or crowdedness of two species in their resource utilization or determining the similarity of two different radiologists' readings of the same x-rays. One such similarity measure, also known as overlap coefficient, has been used for comparing fits of statistical distributions by different methods. The sampling distributions of some measures are estimated using simulations whose parameters are estimated using multiple regression techniques with transformations.

Determinants of High Crude Oil Price: A Nonstationary Extreme Value Approach
Asim Kumer Dey
Princeton University and UT Dallas

There are a number of factors that influence crude oil prices. We use a nonstationary extreme value approach to model crude oil prices. The model includes world oil demand, supply, dollar index, oil consumption, geopolitical and economic events as exogenous factors influencing crude oil prices. We assess the change in crude oil price dynamics as the consequence of changes in covariates. The primary objectives of the study are as follows. Firstly, to investigate the impact of some exogenous variables on high crude oil prices; secondly, to fit the best nonstationary generalized extreme value model; thirdly, to compute the return level of crude oil prices for certain values of exogenous variables; fourthly, to compute the probability of crude oil prices which are extreme in nature; and finally, to measure the relative risk of a high oil price due to different covariate values. The finding shows that it is reasonable to model extreme oil prices using the nonstationary extreme value approach as opposed to the stationary approach. Moreover, the study shows that some of the independent variables considered here have significant roles in high crude oil prices and their changes create high risk of occurring extreme crude oil prices.
This work is joint with Audrene Tiakor (Lamar University) and Kumer Pial Das (UL Lafayette).

Smoothed LSDV Estimation of Semiparametric Functional-Coefficient Panel Data Models with Two-Way Fixed Effects
Shaymal C. Halder
Department of Mathematics and Statistics, Auburn University

This paper considers the estimation of a semiparametric varying-coefficient panel data model with both the individual and time fixed effects. We extend Sun et al.(2009) estimator to also accommodate unobservable unit-invariant common effects that may correlate with explanatory variables in an arbitrary way. The unknown functional coefficients of interest are estimated by (asymptotically) concentrating out both fixed effects via local-linear kernel-smoothed least squares dummy variable approach. Monte Carlo simulations under difference scenarios show that the proposed estimation procedure exhibits good finite-sample performance.
This work is joint with Emir Malikov

Improving Parameters Estimation of Integral Projection Model in Fluctuating Environments

Gopal Nath
Department of Mathematics and Statistics, Auburn University

The most commonly used data-driven models for population dynamics are matrix projection models (MPM), which project discrete population structure (age or size classes) in discrete time. These models are well understood mathematically and there is a well-developed toolbox of techniques for their analysis. In many populations, a continuous trait such as body mass is a key determinant of performance: all else being equal, larger individuals tend to exhibit greater survival and fecundity, so using a continuous state variable will be more appropriate and often improve the performance of the model. Easterling, Ellner and Dixon (2000) originally proposed the integral projection model (IPM) as an alternative to matrix projection model for populations in which demographic rates are primarily influenced by a continuously varying measure of individual size or state. In our study we will analyze how robustness in modeling the continuous size variable affects population growth rate.

Comparing Multiple Linear Regression and Principal Components Regression

G M Toufiqul Hoque
Mathematics Department, Lamar University

The purpose of this study is to compare two modeling techniques, which are multiple regression analysis and principal component regression (PCR) in terms of finding their coefficients and prediction accuracy. In multiple regression, multiple explanatory variables are used to predict the outcome of a response variable. Moreover, the goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. PCR is also a regression analysis technique based on principal component analysis. But in PCR, unlike multiple regression, principal components of the explanatory variables are used as regressors. In this study, air pollution and mortality data is used to measure the negative long-term effect on health because of excessive air pollution. In this data, fifteen different variables of sixty metropolitan statistical areas are used as explanatory variables to predict the total age adjusted mortality rate in the metropolitan area. In these two methods, the coefficient of determination remains same. But in PCR, after discarding the components that do not explain much variance, the coefficient of determination starts decreasing. Since principle components give insight into which linear combinations of variables are responsible for the collinearities, dropping the components cause the coefficient of determination. Importantly, the reason for less coefficient of determination might be the absence of perfect multicollinearity in the original data set.