Eighth Annual Red River Valley Statistical Conference

North Dakota State University
Department of Statistics

Wednesday, May 2, 2018

Eighth Annual Red River Valley Statistical Conference

Session 1: Chair: Ronald Degges Location: Arikawa Room

10:15-10:30 am	Lincoln Larson	Investigating Statistical vs. Practical Significance of the Kolmogorov-Smirnov Two-Sample Test Using Power Simulations and Resampling
10:30-10:45 am	Ying Lin	Testing Parallelism for the Four-Parameter Logistic Model with D-Optimal Design
10:45-11:00 am	Muhammed Saho	A Visualization Method for Course Evaluations and Other Likert Scale Data

Session 2: Chair: Gang Shen Location: Room of Nations

11:15-11:30 am	Di Gao	Bayesian Lasso Models - With Application to Sports Data
11:30-11:45 am	Xiyuan Liu	A Big Data Approach: Conditional Random Field and Gradient Descent
11:45 am-12:00 pm	Yu Sun	Investment Behavior Analysis Based on Tail Risk Management
12:00-12:15 pm	Kekoura Sakouvogui	Does the Type of the Inefficiency Distributions Matter? An Acceptance-Rejection Monte Carlo Simulation of the Stochastic Frontier Analysis (SFA) Models

Session 3a: Poster Session (with Refreshments) Location: Arikawa Room

12:15-12:45 pm	Jace Duffield	Analysis of Wave Propagation through Tomato Plants
	Brooks Flyberg	Analysis of Popcorn Brands
	Thomas Ian Howden	Effect of Nature and Extent of Functional Group Modification on Properties of Thermosets from Methacrylated Epoxidized Sucrose Soyate
	Ashlee Leech	Testing the Difference Absorbencies Between Brands of Paper Towels
	Reid Leighton	Evaluating the Impact of Cooking Methods and Brand on Pasta Cooking Time
	Emily Leonard	Factorial Modelling of Effects of Age and Material on Connector Conductivity
	Joseph Willert	Effect of Surface Type and Ball Size on Bouncy Balls

Session 3b: Poster Session (with Refreshments) Location: Arikawa Room

12:45-1:15 pm	Morgan Berg	Mental Stress Management: Analysis of Guided Imagery and Its Effect
	Kinza Faiyaz	Analysis of Learning Styles
	Kevin Folman	Analysis of Wave Propagation through Perennial Plants
	Dong Geon Han	Probability of Being The Next Babe Ruth
	Michael Heim	The Effect of Water Amount and Fertilizer on Marigold Plant Growth
	Matthew Hilgers	Analysis of the Effects of Vitamin C Dosage and Delivery Method in Odontoblast Length of Guinea Pigs
	Madison Mathiason	Comparing the Effectiveness of Different Stain Removers in a Randomized Block Design

Session 4: Chair: Rhonda Magel Location: Room of Nations

1:15-1:30	Katie Neset	Comparative Analysis of Traditional and Modified DECODE Method is Small Sample Gene Expression Experiments
1:30-1:45	Jingjun Zhao	Differential Gene Co-expression Networks - Bayesian Biclustering Approach
1:45-2:00	Minglian Lin	Nonparametric Bayesian in Genotype Calling for Tetraploidy
2:00-2:15	Xiaoyi Zhou	Exploring Associations between Lifestyles and Metabolic Syndrome in Middle-Age Chinese Population

ABSTRACTS FOR SPEAKERS

(in alphabetical order by first author's last name)

Author: Di Gao

Title: Bayesian Lasso Models – With Application to Sports Data

Dimension reduction is necessary when the number of observations is less than the number of parameters or when the design matrix is non-full rank. Due to Least Absolute Shrinkage and Selection Operator (Lasso)'s desired geometric property, the Lasso method provides a sharp power in selecting significant explanatory variables and has become very popular in dimension reduction in the last 20 years.

This work studied Bayesian Lasso generalized linear models. A hybrid model estimation approach of full and Empirical Bayesian was proposed. A simple and efficient method in the EM step, which does not require sample mean from the random samples, was also proposed. The expectation step was reduced to derive the theoretical expectation directly from the conditional marginal. The findings of this work suggest that future application will significantly cut down the computation load. A simulation study was conducted to test the accuracy and the variation of the estimates. For an application of the Bayesian Lasso Probit Linear Regression to live data, NCAA March Madness (Men's Basketball Division I) was considered. In the end, the predicting bracket was used to compare with the real tournament result, and the model performance was evaluated by bracket scoring system (Shen et al. 2015).

Author: Lincoln Larson

Title: Investigating Statistical vs. Practical Significance of the Kolmogorov-Smirnov Two-Sample Test Using Simulations and Resampling

This research looks at the power of the Kolmogorov-Smirnov two-sample test. The motivation for this research is a large data set containing soil salinity values. One problem encountered was that the power of the Kolmogorov-Smirnov two-sample test became extremely high due to the large sample size. This extreme power resulted in statistically significant differences between two distributions when no practically significant difference was present. This research used resampling procedures to create simulated null distributions for the test statistic. These null distributions were used to obtain power approximations for the Kolmogorov-Smirnov tests under differing effect sizes. The research shows that the power of the Kolmogorov-Smirnov test can become very large in cases of large sample sizes.

Author: Minglian Lin

Title: Nonparametric Bayesian in Genotype Calling for Tetraploidy

Tetraploidy describes one cell contains four sets of chromosome where the additional chromosomes change a range of traits such as decreased growth rate, reduced fertility, gigas effect and so on. But current statistical models describe the allele signals by continuous distribution. Here, I developed a mixture model to do the genotype calling based on discrete data of tetraploidy using nonparametric Bayesian approach.

Author: Ying Lin

Title: Testing Parallelism for the Four-Parameter Logistic Model with D-Optimal Design

It is often important to test for parallelism between a pair of dose-response curves of reference standard and test sample, in order to determine the potency of the test preparation relative to the standard preparation. How to design the testing process is undoubtedly the key factor that determines the success of the experiment. Compared with classical designs, optimal design is known to be more powerful and particularly useful. It’s a straight optimization based on a chosen optimality criterion and the model that will be fit. In this study, D-optimal design is implemented to study the parallelism and compare its performance with some known standard classical design. To compare the power of standard classical design and D-optimal design, I modified D-optimal design to test the parallelism in the four-parameter logistic (4PL) response curves using Intersection-Union Test (IUT). IUT method is appropriate when the null hypothesis is expressed as a union of sets, and by using this method complicated tests involving several parameters are easily constructed. Since D-optimal design minimizes the variances of model parameter estimates in same sense, it can bring more power to the IUT test. A simulation study will be presented to compare the empirical properties of the two different testing designs.

Author: Xiyuan Liu

Title: A Big Data Approach: Conditional Random Field and Gradient Descent

With the progressing of our technology, collecting data become much easier. However, that also bring up another problem: How to model and efficiently analyze big data. A novel machine learning approach is Conditional Random Field (CRF), which is a very general and flexible to labeling observations. The CRF is much more powerful when the dimension of its parameter space is very large, however, it in return, renders model estimation of CRF quite challenging: The traditional

Newton-Raphson computation method fails because of large dimension of Hessian matrix. In this presentation, I will propose how to estimate 100 dimensional CRF using Gradient Descent and present the efficiency of CRF via a simulation data.

Author: Katie Neset

Title: Comparative Analysis of Traditional and Modified Decode Method in Small Sample Gene Expression Experiments

Background: The DECODE method integrates differential co-expression and differential expression analysis methods to better understand biological functions of genes and their associations with disease. The DECODE method originally was designed to analyze large sample gene expression experiments, however most gene expression experiments consist of small sample sizes. This paper proposes modified test statistic to replace the traditional test statistic in the DECODE method. Using three simulations studies, we compare the performances of the modified and traditional DECODE methods using measures of sensitivity, positive predictive value (PPV), false discovery rate (FDR), and overall error rate for genes found to be highly differentially expressed and highly differentially co-expressed.
Results: In comparison of sensitivity and PPV a minor increase is seen when using modified DECODE method along with minor decrease in FDR and overall error rate. Thus, a recommendation is made to use the modified DECODE method with small sample sizes.

Author: Muhammed Saho

Title: A Visualization Method for Course Evaluations and other Likert Scale Data

Student rating of instruction (also known as course evaluation) is one of the primary ways of collecting feedback from students at NDSU. Since almost every student in every course submits one at the end of the semester, it generates a lot of data. The data is summarized into text based reports with emphasis on average rating of each question. At one page per course, analyzing these reports can be overwhelming. Furthermore, it is very difficult to identify patterns in the text reports.

We combine heat maps and small multiples to introduce a visualization of the data that allows for easier comparison between courses, departments, etc. Our research consisted of three main components. First, we defined a data format for storing and transmitting the raw data. Next we built an interactive web application that consumes the aforementioned data format and generates the visualizations. Finally, we simulated reference data to facilitate interpretation of the visualizations. Additionally, we discussed how our research can be applied more generally to Likert scale data.

Author: Kekoura Sakouvogui

Title: Does the Type of the Inefficiency Distributions Matter? An Acceptance-Rejection Monte Carlo Simulation of the Stochastic Frontier Analysis (SFA) Models

The Stochastic Frontier Analysis (SFA) is widely used to estimate the individual efficiency measures in the banking industry. A major criticism related with the application of SFA of Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977) is the lack of justifications for the choice of the inefficiency distribution. Because SFA yields different efficiency scores based on the chosen inefficiency distributions, researchers and policy makers always face the problem of determining the true efficiency of banks (Ander and Hesse, 2011). Earlier researchers have not addressed the issues of inefficiency distributions. To allow for comparative technical efficiency measures across different inefficiency distributions (truncated normal, half normal, and exponential), we provide a Monte Carlo simulation under the acceptance- rejection method of comparative scale to the inefficiency distributions. Through extensive simulation designs, the results show that when the inefficiency distributions are on the same scale, researchers and policy makers should always choose the truncated normal SFA model.

Author: Yu Sun

Title: Investment Behavior Analysis Based on Tail Risk Management

Traditional portfolio theories typically target rational investors who pursue optimal portfolio investment that balances the trade-off between risk and return. However, a more comprehensive study should incorporate investors’ psychological factors, especially those of pessimistic and irrational investors’. This paper is to model investors’ investment behaviors in response to the market frictions by managing the tail risk of asset portfolio. We employ the Conditional Value-at-Risk(CVaR) as the measure of tail risk. Specifically, we build four models that respectively (1) maximizes the left-tail CVaR, (2) minimizes the right-tail CVaR, (3) minimizes the left-tail CVaR, and (4) jointly maximizes the left-tail CVaR & minimizes the right-tail CVaR. We contribute to the literature by applying portfolio optimization approaches to analyze various groups of rational and irrational investors. Through portfolio optimization models that manage portfolio tail risk, we investigate behavioral-based investment strategies to accommodate investors’ needs/concerns. This paper also sheds light on exploring economic information regarding asset pricing puzzles and long-run risk based on the investment behaviors of traditional and non-traditional investors.

Author: Jingjun Zhao

Title: Differential Gene Co-Expression Networks – Bayesian Biclutering Approach

Gene co-expression networks constructed from microarray data or RNA sequencing data can be used to group genes of unknown biological functions, and differential gene co-expression networks can be further used to rank candidate disease genes, to recognize transcriptional regulatory programs, or to implement other important gene analysis. An explicit gene co-expression network recovered from high-throughput experiments to measure gene expression levels can notably enhance the interpretation of the networks. Biclustering, a data mining technique which allows simultaneous clustering of the rows and columns of a matrix, has been adopted in construction of gene co-expression networks. Here, I review a robust and computationally tractable Bayesian statistical model for biclustering genes to recover non-disjoint clusters of co-expressed genes in subsets of samples using gene expression level data. The recovered clusters are used to build gene co-expression networks jointly across all genes by computing the full regularized covariance matrix between all pairs of genes instead of testing each possible edge separately. Compared with other related biclustering methods, this model recovers latent structure with higher precision across diverse simulation scenarios.

Author: Xiaoyi Zhou

Title: Exploring Associations between Lifestyles and Metabolic Syndrome in Middle-Age Chinese Population

Nowadays the prevalence of Metabolic Syndrome (MetS) affects many middle-age people in China. MetS is associated with the risk of type 2 diabetes and cardiovascular disease. Identifying the potential risk factors contribute to MetS is very important for preventing cardiovascular disease. The associations between lifestyles and prevalence of MetS are extensively studied by researchers. A cross-sectional study, which was conducted by Strand, MA. surveyed 659 subjects in Yuci, China in 2012. The proportional odds model was applied to determine the associations between lifestyles and MetS in three Chinese middle-age groups. The results demonstrated that doing daily exercise was one of the best method to treat MetS. Moderate alcohol consumption could prevent MetS in age group born in 1956. Occasionally milk consumption could prevent MetS in age group born in 1964, while it did not help age groups born in 1960-1961 and in 1956.

ABSTRACTS FOR POSTER PRESENTERS

(in alphabetical order by first author's last name)

Author: Morgan Berg

Title: Mental Stress Management: Analysis of Guided Imagery and Its Effect

Patients may often face feelings of apprehension and unease before undergoing major surgeries or medical procedures. The use of guided imagery is becoming more prevalent in the medical field to reduce anxiety and fear in patients. Guided imagery is a therapeutic technique used to produce a more positive scenario that is now being used before patients proceed with the medical procedure. In this Randomized Complete Block Design experiment, guided imagery will be used on patients before surgeries to test if there is a difference in pulse rate while taking into consideration factors such as age and gender (male or female). To statistically analyze the results, an ANOVA will be used to examine the effectiveness of guided imagery on pulse rate.

Author: Jace Duffield

Title: Analysis of Wave Propagation through Tomato Plants

With the increase use of wireless soil sensors in precision agriculture, a study must be performed to determine various effects on signal strength. The purpose of this study was to analyze the effect of plant growth on the signal strength of an embedded wireless soil sensor. In this study, we used a two horn antenna system to study the signal propagation with respect to plant growth using a test bed containing tomato plants. The results showed that there were no differences in signal strength with regards to the plant growth. The results from this study will benefit the design process of future wireless soil sensor systems as the agriculture business moves towards higher precision.

Author: Kinza Faiyaz

Title: Analysis of Learning Styles

This study investigates how the learning styles of student vary depending on their gender (Male or Female) and academic level (undergraduate or graduate). Students were given a 40-word pair questionnaire in which they had to decide which of the two words is more characteristics of their learning style when compared to other. They could describe the chosen characteristic by options like, most of the time, over half of the time, and about half the time. Numerical value is associated with each of the options, so it would be easier to add all of their scores. Learning Style Questionnaire contains four scales Abstract Conceptualization (AC), Abstract Experimentation (AE). Concrete Experience (CE), and Reflective Observation (RO). Statistical Analysis will be performed on the results collected from this questionnaire. ANOVA will help in determining the statistical difference of learning styles between gender and academic level.

Author: Kevin Folman

Title: Analysis of Wave Propagation through Perennial Plants

In precision agriculture, the use of wireless soil sensors is on the rise. With this increase, a study must be conducted to pinpoint the possible effects on signal strength. In one soil bed separated into three different plots: a control plot, a small perennial plot, and a large perennial plot, a study was created to test signal strength versus the growth of plants. With an embedded wireless soil sensor and a two-horn antenna system, the wave propagation was measured throughout the growing season of the perennials. The results from this study showed that there were no significant differences in the signal strength when the plants grew. However, the results from this study will still benefit the future development of wireless soil sensors.

Author: Brooks Flyberg

Title: Analysis of Popcorn Brands

This presentation will compare the number of kernels left not popped across three different brands of popcorn. Each batch will be 400 kernels which is roughly the amount of a single bag of microwavable popcorn. Each trial will be given 3 replicates at 3 different levels of oil for each brand. This test will try to distinguish the better brand and maybe optimize the amount of oil that should be used to make the popcorn. My results will determine if the brands and amount of oil are significantly different. The results of the test could also be compared to the data of each popcorn brands popping percentage given on their websites.

Author: Dong Geon Han

Title: Probability of Being the Next Babe Ruth

This year, a baseball player whose name is Shohei Ohtani from Japan came to Major League Baseball. The special thing is that he is a two-way player, which means he is a pitcher as well as a hitter. Babe Ruth was the only one two-way player who won 10 games and hit 10 home runs in one season 97 years ago. I wondered if Ohtani will accomplish 10-10. So, I compared Japanese pitchers' records, which are Earned Run Average (ERA) and average win per year, in Japanese Baseball League (NPB) with the pitchers' record in MLB to predict the probability of 10 wins. Also, I compared Japanese hitters' record, batting average (AVG) and average home run per year, in NPB with the hitters' record in MLB to predict the probability of 10 home runs. Furthermore, I predicted how many wins and home runs he would get at the end of the season based on average fastball speed and home runs divided by fly balls. I used the record of MLB players who have the similar physical and statistical condition to him for comparison.

Author(s): Thomas Ian Howden (Arvin Z. Yu, Jonas M. Sahouani, Raul A. Setien, Dean C. Webster)
Title: Effect of Nature and Extent of Functional Group Modification on Properties of Thermosets from Methacrylated Epoxidized Sucrose Soyate
A study was conducted by researchers in NDSU's Coatings and Polymeric Materials department to evaluate the impact of modification of epoxidized sucrose soyate on the properties of the resins and thermosets. Soyates were modified by using a combination of methacrylate and acetic, propionic, and butyric inert esters. Three variable were analyzed in this experiment; glass transition temperature, mechanical properties, and resign viscosity. For the purpose of this project, I focus on only the glass transition temperature variable. The objective was to maintain high transition temperatures and good mechanical properties while obtaining a low viscosity by replacing methacrylate groups with inert esters. Transition temperatures of resigns were analyzed at increasing amounts of styrene and free radically cured using peroxyesters as initiators. The thermosets produced from this process had improved flexibility and toughness with only a slight reduction in the glass transition temperature. These bio-based resigns create numerous new end-use applications.

Author: Michael Heim
Title: The Effect of Water Amount and Fertilizer on Marigold Plant Growth
Fertilizer is commonly used in commercial agriculture and residential horticulture. This study used Marigolds, a low maintenance plant, to see if different brands of garden fertilizers produce better results. Similarly, the plants were given different water amounts to see the effect of overwatering and lack of water. The plants' heights were measured and analyzed during this experiment to determine the ideal water amount and if a certain brand of fertilizer produced the tallest plants. Since the fertilizers used in this experiment were water soluble, the interaction between water amount and fertilizer, i.e. the concentration of the mixture, was also analyzed.

Author: Matthew Hilgers

Title: Analysis of the Effects of Vitamin C Dosage and Delivery Method in Odontoblast Length of Guinea Pigs

This presentation will compare the length of odontoblasts in guinea pigs at different levels of dosage and delivery method. There were three levels of vitamin C dosages ranging from .5 to 2 mg/day. The second factor was delivery methods of which there where two; orange juice and ascorbic acid. An analysis of variance will be done to determine if either factor had an effect on the response variable as well as the two way interaction effect.

Author: Ashlee Leech

Title: Testing the Difference Absorbencies between Brands of Paper Towels

This experiment compares the different absorbencies between different brands of paper towels and different amounts of water and juice. There will be 3 different brands of paper towels with 3 different amounts of liquid. When poured from the same height, the area of wetness on the paper towel will be calculated. A two-stage nested design with the two-stage nested effect model will be used. An ANOVA table will be used to determine the statistical inferences to see if there is a difference in absorbencies between the brands or a difference in absorbencies between liquids.

Author: Reid Leighton

Title: Evaluating the Impact of Cooking Methods and Brand on Pasta Cooking Time

Many college students don't have enough time for cooking, often balancing both work and class. One of the simplest meals to make is pasta, but what is the fastest way to make such a dish? This experiment focuses on two factors. The first compares how quickly three different brands of pasta cook. The second compares adding water and pasta at the same time against boiling water and then adding the pasta. This is a full factorial effect model with interaction to see how different methods of cooking effect different brands.

Author: Emily Leonard

Title: Factorial Modelling of Effects of Age and Material on Connector Conductivity

Metals are known to be susceptible to rust and corrosion, so older metals may not perform as well as newer pieces. As such, it is worth investigating the effect of material age on conductivity. This experiment aims to analyze the effect of two factors, elapsed days and connector type, on voltage loss in an electrical circuit. Using a completely randomized design, electrical circuits were tested in randomly and equally assigned treatment groups covering a variety of exposure times and material types. The response variable, loss of voltage, is measured for each circuit. A full factorial model was used to test the main and interaction effects of connector age and connector type to compare the connector types to discern which is the most robust.

Author: Madison Mathiason

Title: Comparing the Effectiveness of Different Stain Removers in a Randomized Block Design

The effectiveness of three stain removers is tested on four types of stains in a completely randomized block design. Each stain remover, Shout, OxyClean, and Espro Sports Cleaner is applied on four stains: pasta sauce, mud, coffee, and red wine. The appearance of the stain is rated on a 1-100 scale (with 100 being a completely removed stain) after the stain remover is applied. The data is modeled in an additive factorial effect model and analysis of variance (ANOVA) is used.

Author: Joe Willert

Title: Effect of Surface Type and Ball Size on Bouncy Balls

Bouncy balls are used by kids around the world for a source of entertainment. In this study, the effect of surface type and the size of a ball on the initial bounce of a bouncy ball was examined in order to determine the most effective combination of factors to generate the most excitement. The study looked at three different types of surfaces commonly found in households today: tile, short-haired carpet, and wood. The study also looked at the use of small vs large balls both made of the same material. Taking these two factors into account, one is able to find the best combination in order to generate the max amount of fun for all ages.