Sixth Annual Red River Valley Statistical Conference

North Dakota State University
Department of Statistics

Friday, April 29, 2016

Sixth Annual Red River Valley Statistical Conference

Session 1: Chair: Rhonda Magel Location: Meadow Room

10:05 am *Predicting Winners of NCAA Women's Basketball Tournament Games,* Wenting Wang
10:20 am *Modeling the Winners of NCAA Women's Division II Basketball Tournament Games,* Feifei Huang
10:35 am *Prediction on 2015-2016 Season NCAA March Madness Bracket -- Based on Bayesian Estimation*, Di Gao
10:50 am *Prediction on 2015-2016 Season NCAA March Madness Bracket -- By Support Vector Machine*, Qian Wen
11:05 am *Predicting Outcomes of NBA Basketball Games*, Scot Jones
11:20 am *Point Spread for Women's Volleyball*, Deling Zhang
11:35 am *Comparative Classification of Prostate Cancer Data Using the Support Vector Machines, Random Forest, DualKS and k-Nearest Neighbors*, Kekoura Sakouvogui

Session 2: Chair: Gang Shen Location: Lark Room

10:05 am *How to value the Major League Baseball Player*, Anthony Kopka
10:17 am *Analysis of Effects of Acid Scarification for Seeds*, Cole Bishop
10:29 am *Analysis of Honey Locust Seed Germination,* Austin Todd & Cory Schwaab
10:41 am *Analysis of Age Difference between Eye-Tracking: Motion and Motion Parallax,* Colton Ulmer, Kinza Faiyaz & Lauren Tupa
10:53 am *Experimental Design for Determining Which Species of the Wheat Can Be Well Disease-Resistant,* Jianfeng Zhang
11:05 am *D-optimal Design for the 5PL-1P Model in Chemical Toxicity Assessment,* Jenna MacDonald
11:20 am *Adaptive c-Optimal Designs for Estimating the Edp,* Anqing Zhang
11:35 am *The Communication of Discrete Emotions via Instrumental Musical Timbre,* Ezekiel Brockmann

Session 3 (Poster Session with Refreshments) Location: Peace Garden Room

11:50 am – 12:30 pm
Information Asymmetry in Budget Allocation: Analysis of the Truth-inducing Incentive Scheme, Yun Zhou

Session 4: Chair: Yarong Yang Location: Room of Nations

12:30 pm *Undergraduate Collaboration: Using Learning Assistants in a Large Statistics Class*, Ronald Degges (Keynote Speaker)
1:00 pm *Identification of Differentially Expressed Genes using DESeq2,* Ekua Kotoka
1:15 pm *A Study Case on Comparison between Microarray Gene Expression and RNA-seq*, Qi Wang
1:30 pm *Establishing a Threshold to Determine Whether the Gene in the Plant is Changed by Virus Using Hypothesis Test*, Xiyuan Liu
1:45 pm *Analyzing Borderlands 2's Loot System: Parcel and Item Drop Rates*, Kathryn Campbell

ABSTRACTS FOR SPEAKERS

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

Author: Cole Bishop

Title: Analysis of Effects of Acid Scarification for Seeds

The use of seed scarification is vital for any grower that wishes to circumvent the time needed seeds to overcome natural physical dormancy. Many seeds are receptive to different methods of overcoming physical dormancy, such as mechanical scarification (removal of coat with knife or other devices, etc.) exposure to boiling water and other chemical acids (household bleach or acids). From a pilot study, it was observed that exposure to sulfuric acid delivers the strongest germination frequency and seed health. My experiment therefore exposed seeds to various levels of sulfuric acids and observed their development and germination proportion.

Author: Ezekiel Brockmann

Title: The Communication of Discrete Emotions via Instrumental Musical Timbre

Music is an excellent tool through which people communicate emotional experience. This study is built upon the Brunswikian Lens Model adapted specifically for musical communication. Compared to the massive amount of research on emotional correlation to higher order musical structures such as mode, melody, phrase, genre and so forth, relatively little has been investigated with regard to the role of musical timbre in the communication of emotional intention. It has been shown that listeners consistently attribute emotional dimensions to isolated instrument samples but novel in this study is how musical performers manipulate their instruments to create five discrete emotions: Happy, Sad, Angry, Fearful, and Tender. Amateur musicians gave permission to be recorded performing the basic 5 discrete emotions using only one short note per emotion, manipulating whatever other elements of sound they deem necessary in order to effectively communicate a particular emotion. The recorded timbral stimuli then undergo two challenging assessments. They are presented to a group of 32 listeners in a perceptual experiment to discover if the recognition of the emotional content of the timbral signals is possible and that the Brunswikian Communication model can be maintained under these stringent limitations. The second assessment of the performances is computational, and a brief explaination is given about of the analysis of the timbral recordings using the Music Information Retrieval (MIR) Toolbox in the MatLab. A veritable "unwelcome bounty" of variables emerges from this process and the problem of applying appropriate methods of data reduction to the psychoacoustic dataset is addressed. It is shown that communication of five discrete emotions is possible, even as the process of teasing out and adequately

"discovering" the specific psycho-acoustic channels through which the communication of emotional information occurs is a much more complex process.

Author: Kathryn Campbell

Title: Analyzing Borderlands 2's Loot System: Parcel and Item Drop Rates

Borderlands 2 is a video game that uses the accumulation of goods, referred to in game as loot, to advance a player's character and the game's plot. The game uses a procedural random number generator with item pools as the foundation for its loot system. The creators of *Borderlands 2* claim that, unless specified in game, a lootable object is independent from other objects and has identical item probabilities within its object class. A player has reported a parcel, a class of lootable object, with an above average rate of containing a weapon. The purpose of this project is to test the validity of this claim by sampling parcels on a single map and then applying Fisher's exact test. The analysis showed that of the sampled parcels their respective locations had no effect on the probability of a weapon appearing.

Author: Ronald Degges (Keynote Speaker)

Title: Undergraduate Collaboration: Using Learning Assistants in a Large Statistics Class

Undergraduate learning assistants in statistics courses improve student performance and retention. Engaging students with real life applications in statistics during class time motivates them to learn more statistical techniques and processes. The integration of experiential learning across two years will be examined through the implementation of a Learning Assistant (LA) Program. Test scores have shown consistent improvement in student performance and retention rates were maintained as LAs engaged students with various in-class statistics problems. Results suggest that students need at least 1 LA to 40 students to make the experiential learning process most successful.

Author: Di Gao (under direction of Rhonda Magel and Gang Shen)

Title: Prediction on 2015-2016 Season NCAA March Madness Bracket -- Based on Bayesian Estimation

NCAA's March Madness is becoming more and more popular. Based on AGA's (American Gaming Association) report (2015), about 40 million people filled out 70 million March Madness brackets (Moyer, 2015). Every participant wants to get correct predictions of game results. This paper introduced a prediction method, which involves Bayesian estimation along with a generalized liner model with logit link (Probability Self-Consistency Model). This allows us to borrow the information of historical

winning rate between teams of different seeds. In addition to this, a SIR (Simple Importance Resampling) algorithm is used to estimate the covariates' coefficient. Finally, based on the estimated coefficients, the prediction bracket can be developed; and the prediction accuracy can be calculated to evaluate the model.

Author: Feifei Huang

Title: Modeling the Winners of NCAA Women's Division II Basketball Tournament Games

This presentation first presents a least squares regression model to identify the in-game statistics that help explain the variation in point spread for an NCAA Division II Women's Basketball Tournament games. Then a logistic regression model is presented to estimate the probability of a team winning a tournament game based on the differences in significant in-game statistics. Both models are developed based on the tournament in-game statistics in years 2012, 2013, and 2014. Differences in the following variables are significant in both models: field goal percentage, 3-point field goal percentage, free throw percentage, offensive rebounds, personal fouls and turnovers. Difference in assists is only significant in the point spread model. Both models are validated using the in-game statistics for the 2015 tournament, indicating a prediction accuracy as high as 95.24%. Seasonal averages for the 2014 – 2015 season are then used to predict game results in the 2015 tournament. The prediction accuracies are 60.32% and 66.67% for the point spread model and the logistic regression model, respectively.

Author: Scot Jones

Title: Predicting Outcomes of NBA Basketball Games

A stratified random sample of 144 NBA basketball games was taken over a three-year period, between 2008 and 2011. Models were developed to predict point spread and to estimate the probability of a specific team winning based on various in-game statistics. The year the games were played did not matter. The models were verified using exact in-game statistics for a random sample of 50 NBA games taken during the 2011-2012 season, and were found to have an accuracy of 94%. Three methods were used in an effort to estimate in-game statistics of a future game so that the models could be used to predict a winner in a game played by Team A and Team B. Models using these methods had accuracies of approximately 62%. The following in-game statistics were significant in the model: field goal shooting percentage, three-point shooting percentage, free throw shooting percentage, offensive rebounds, assists, turnovers, and free throws attempted. Seasonal averages for these in-game statistics

will be found for each team in the playoffs and will be used in the model developed to predict the winner of each game for the 2013, 2014, and 2015 NBA Championships.

Author: Anthony Kopka

Title: How to value the Major League Baseball Player

Baseball players are well known for making an extraordinary amounts of money, like Giancarlo Stanton who stands to make $325 million dollars over the next 12 years, averaging $25 million a year. This study was conducted to determine the regression analysis of a baseball players worth based on their offensive statistics. One of the goals is to find the prediction/regression equation to determine how much a player should be paid based on their offensive categories such as the amount of games played in a season, runs scored, total bases, home runs, runs batted in, walks, strikeouts, stolen bases, batting average, on base percentage, and slugging percentage. My methodology of collecting my data was using the official website of major league baseball (mlb.com) to find the statistics and salaries of the 300 players over the last 4 seasons. Then I used Minitab software to analyze the 1200 lines of data and draw my conclusions. Some variables that I didn't account for that swayed some of my results was injuries and players not living up to large and long contracts that they had signed. I still haven't drawn up my final conclusions and results.

Author: Ekua Kotoka

Title: Identification of Differentially Expressed Genes using DESeq2

Next generation sequencing technology, RNA-Sequencing (RNA-Seq), is becoming the preferred approach over the traditional microarrays for characterizing and quantifying entire transcriptomes. RNA-Seq provides quantification of gene expression using counts of reads recorded to a particular gene, whereas array-based technologies measures intensities using continuous distributions. Using RNA-Seq to identify differentially expressed (DE) genes depends on estimating the variability of the read-counts. These estimates are mostly based on statistical models such as the negative binomial distribution to address the overdispersion problem. We present the DESeq2 method used for differential analysis of count data; this method uses shrinkage estimation of dispersions and fold changes to improve stability and interpretability of estimates. Finally, we illustrate the use of the DESeq2 method by analyzing C57BL/6J (B6) and DBA/2J (D2) mouse strains samples.

Author: Xiyuan Liu

Title: Establishing a Threshold to Determine Whether the Gene in the Plant is Changed by Virus Using Hypothesis Test

Plant scientists who analyze the gene activities in the infected plant are often faced with one question: How to determine whether the gene in the plant is changed by the virus. This study developed a method, which establishes a threshold, to answer the question. The threshold is based on a hypothesis test and only genes that changed by the virus can pass it. In addition, the study solved the problem of conducting a hypothesis test with-out knowing the distribution under the null hypothesis. This study also introduced Benjamin-Hodge method to constrain the family-wise type one error. The study can be useful to any scientist who wants to separate a set of subjects into two di□erent subsets.

Author: Jenna MacDonald

Title: D-optimal Design for the 5PL-1P Model in Chemical Toxicity Assessment

The five-parameter logistic minus one-parameter model is a hybrid between the five-parameter model and the four-parameter model used for the relationship between dose concentration and response. The four-parameter model includes the maximum concentration, minimum concentration, slope, and the median concentration level EC_{50} as its four parameters. The five-parameter model add an asymmetric factor which is important due to deviations from the symmetry of the sigmoid curve. The five-parameter model however is more difficult to fit due to the addition of the fifth parameter which is why the 5PL-1P model is used so that the asymmetric factor is taken into account but it has less parameters to fit so it is easier to work with. For the 5PL-1P model, D-optimal designs are obtained to estimate the model parameters effectively which lead us to estimate the relationship between the response and the concentration the most accurately. Then we compare the D-optimal designs to the designs that are used to study the 5PL-1P model in real toxicity assessment. We show that the D-optimal designs works better than the real designs by comparing their efficiencies and comparing mean squared errors through simulation studies.

Author: Kekoura Sakouvogui

Title: Comparative Classification of Prostate Cancer Data Using the Support Vector Machines, Random Forest, DualKS and k-Nearest Neighbors.

This paper compares four classifications tools, Support Vector Machine (SVM), Random Forest (RF), DualKS and the *k*-Nearest Neighbors (*k*NN) that are based on different statistical learning theories.

The dataset used is a microarray gene expression of 596 male patients with prostate cancer. After treatment, the patients were classified into one group of phenotype with three levels: PSA (Prostate-Specific Antigen), Systematic and NED (No Evidence of Disease). The purpose of this research is to determine the performance rate of each classifier by selecting the optimal kernels and parameters that give the best prediction rate of the phenotype. The paper begins with the discussion of previous implementations of the tools and their mathematical theories. The results showed that three classifiers achieved a comparable performance that was above the average while DualKS did not. We also observed that SVM outperformed the *k*NN, RF and DualKS classifiers.

Authors: Austin Todd & Corey Schwaab

Title: Analysis of Honey Locust Seed Germination

This presentation will compare the seed germination rates of honey locusts seeds at different levels of acid soak time. Data was collected by students in the plant sciences department from over 1200 planted seeds. There were 12 levels of acid soak time ranging from 0 – 20 hours. We will be using a least squares regression to show the effect of soak time as well as testing the effects of each replicate on germination rates. We will be using multiple comparison tests to evaluate the different levels of soak time as well as testing our variance assumptions. We will also use an ANCOVA model to determine the best soak time and compare both methods. Our results will determine which level of soak time yields the highest rate of germination.

Authors: Colton Ulmer, Kinza Faiyaz & Lauren Tupa

Title: Analysis of Age Difference between Eye-Tracking: Motion and Motion Parallax

An examination of the difference in performance from adults aged between eighteen to twenty-one years old and adults aged at over sixty to seventy-three years old at tracking motion and motion at a depth. Subjects were tasked with determining which direction a target blip would move towards on a computer monitor. The target blip would then either be alone, or have cues of depth to distinguish motion parallax. From there, the target blip either would or would not have a "ghost" image following behind to distinguish between non-tracking and tracking treatments. And finally, the participants would either be allowed or not be allowed to "chase" the target blip with their eyes to distinguish between the pursuit and non-pursuit treatments. This was a Within-Subjects design as every participant attempted each treatment, and to combat ordering bias each participant attempted the tasks in a random order. An analysis of variance was conducted due to differences in sample sizes between age groups.

Statistical analysis was performed through SAS, and from there the subsequent ANOVA table was used to determine the significance and effect size of the treatment factors and factorial effect model.

Authors: Qi Wang, Kurt Zhang & Yarong Yang

Title: A Study Case on Comparison between Microarray Gene Expression and RNA-seq

Microarray and RNA-seq (RNA sequencing) are two commonly used technologies to measure the expression levels of large numbers of genes simultaneously. To study the relationships between the gene expression levels measured by microarray and RNA-seq would help us understand better about the two technologies and hopefully provide more information about the genes. Data transformations, regression models and quantile normalization have been used to reveal possible relationships between the data obtained by microarray and RNA-seq.

Author: Wenting Wang

Title: Predicting Winners of NCAA Women's Basketball Tournament Games

Women's basketball game is becoming more and more popular, spreading from the east coast of the United States to the west coast, in large part among women's colleges. The National Collegiate Athletic Association (NCAA) Women's Division I Basketball Tournament is one of the most famous ones. It is also known as March Madness or the Big Dance. March Madness has ranked itself the second place, following NFL, in sports gambling market in terms of wage. It is the basketball tournament that grabs most eyeballs from the audience because of fun from the viewpoints of both pure sports and gambling. The objectives of this study are: Use ranking difference and Actual score difference to develop least squares regression models to predict terms that win for the variation runs of the NCAA women's basketball game. Statistic data were collected from 2008-2009, 2009-2010, 2010-2011 and 2011-2012 seasons and analyzed using least squares regression methods. Prior game median statistics were collected for teams competing in a sample of games from 2012-2013 and 2013-2014 to determine the accuracy of the models.

Author: Qian Wen (under direction of Rhonda Magel and Gang Shen)

Title: Prediction on 2015-2016 Season NCAA March Madness Bracket – By Support Vector Machine

"March Madness," the term used to describe the excitement created by the National Collegiate Athletic Association (NCAA) Men's Division I Basketball Tournament from the second week of March

through the first week of April. The game is single-elimination tournament played by 68 college basketball teams to determine the national championship of the major college basketball teams. This presentation will focus on bracketing the NCAA Men's Division I Basketball Tournament using Support Vector Machine method (SVM). It is a supervised method that attempts to find a decision boundary between two classes (win and loss) by maximizing the margin between the two classes. The detail about the SVM method and the performance on predicting NCAA Men's Division I Basketball Tournament bracket will be demonstrated in this talk.

Author: Anqing Zhang

Title: Adaptive c-Optimal Designs for Estimating the Ed*p*

The four-parameter logistic model is often used to describe the dose-response relationships in many dose finding trials. Under the four-parameter logistic model, optimal designs to estimate the EDp accurately is studied. The EDp is the dose achieving 100p% of the Emax (the maximum expected response- the minimum expected response). C-optimal design works the best to estimate the EDp, however, it truly depends on the model parameters. In order to avoid the dependence problem, we propose adaptive c-optimal designs for estimating the EDp. A two-stage approach is used to construct the adaptive c-optimal designs.

Author: Deling Zhang

Title: Point Spread for Women's Volleyball

Volleyball has become well-known and competitive sports with physical and technical performance over the years. The game results are determined on some important factors such as players, and the team's skills to succeed in a championship. In this project, we selected four independent variables such as number of kills, Errors, hitting percentage and PCT from 108 games. We proposed to analyze volleyball data by using a multiple regression model and logistic regression model. First, we develop a multiple regression with full model and partial model, then we performed F-test for the goodness fit in the model. Second, we will need to validate the selected regression model by collecting new data and compare the results of the regression model to previous results.

Author: Jianfeng Zhang

Title: Experimental Design for Determining Which Species of the Wheat Can Be Well Disease-Resistant

The plant scientists often face the question that how genes on DNA determine which species of the wheat can be disease-resistant. However, genes at different position may have different expressions. In order to solve the problem, a model from experimental design is applied based on the data which is provided by the Department of Plant Pathology. This model is used to search the species of wheat with high disease-resistance. Furthermore, the significance of this analysis can help the researcher to gain a further understanding of the relationship between the species of wheat and the disease-resistance.

ABSTRACTS FOR POSTERS

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

Author: Yun Zhou

Title: Information Asymmetry in Budget Allocation: Analysis of the Truth-inducing Incentive Scheme

In budget allocation, information asymmetry occurs when the project managers have better information regarding the project costs than the portfolio manager. Information asymmetry between the project managers and the portfolio manager creates an opportunity for the project managers to negotiate for budgetary slack to improve their performance when their compensation is associated with budget attainment, resulting in a false evaluation of managers' budget needs. One way to avoid the problems caused by information asymmetry is the use of a truth-inducing (TI) incentive scheme. The TI incentive scheme is one tool to reduce budgetary slack caused by the project managers - portfolio manager information asymmetry. We identify the value of penalty coefficients in the TI incentive scheme when information asymmetry is present. We first describe the allocation method that achieves budget optimization by assuming that the uncertain costs follow a normal distribution and determine the allocated budget depends on the mean and the standard deviation of the uncertain cost of the project. Then we demonstrate the process of identifying the penalty coefficients based on the judgments of the portfolio manager and the project managers regarding the budget optimization. We

report a lower bound on the ratio between the penalty coefficients in the TI incentive scheme. We conclude that the penalty coefficients for being over budget should be reduced when the portfolio budget is extremely tight, and it should be increased when the portfolio budget is moderate tight.