Second Annual Red River Valley Statistical Conference

North Dakota State University
Department of Statistics

Wednesday, April 18, 2012

Second Annual Red River Valley Statistical Conference

Session 1:	12:30 - 1:30 pm	Loftsgard 380
Break 1:	1:30 - 2:00 pm	Loftsgard 262
Session 2:	2:00 - 3:20 pm	Loftsgard 380
Break 2:	3:20 - 3:50 pm	Loftsgard 262
Session 3:	3:50 - 4:50 pm	Loftsgard 380
Posters:		Loftsgard 260

Session 1: 12:30 pm
An Introduction to Data Analysis in Sports
Kyal Brandt & Joe Long

Should the Giants have taken a knee at the end of Super Bowl XLVI instead of scoring a touchdown? Who is the best defensive shortstop in MLB? Ever since the inception of many sports there has been a desire to quantify a player's performance, or a team's strategy. The sports industry is big business in the United States. Managers, coaches, and players want to make informed decisions involving the game, and as a result, in-depth data analysis has become a very popular topic. *Moneyball,* the book and now movie, has also helped to popularize so called "Sabermetrics." This refers to the development of statistics in baseball used to more accurately reflect in-game activity as compared to traditional statistics. In depth data analysis has spread from its use in baseball and is now done for almost all sports. We will discuss some of the statistics that have been developed. Fans have also shown a strong desire for statistics to enhance the entertainment and understanding of the game. As a result, the Sloan MIT Sports Analytics Conference has become increasingly popular since its inaugural year in 2006. We recently attended this conference and will discuss some of the topics and research that were discussed. One such topic is collecting quality data to analyze which is paramount to conducting any useful analysis. Current technology utilizes real-time video tracking software to efficiently collect data. Other methods, such as utilizing software and tablet applications, are also becoming popular means of collecting data from sporting contests. The availability of such rich data sets makes future data analysis in sports very exciting with the potential to dramatically change the way the games are played.

Session 1: 12:50 pm

Optimal Designs for the Hill Model with Three Parameters

Travis Dockter

Optimal design specifies design points to use and how to distribute subjects over these design points in the most efficient manner. The Hill model with three parameters is often used to describe sigmoid dose response functions. In our paper, we study optimal designs under the Hill model. The first is D-optimal design, which works best to study the model to fit the data. Next is c-optimal design, which works best to study a target dose level, such as ED50 - the dose level with 50% maximum treatment effect. The third is a two-stage optimal design, which considers both D-optimality and c-optimality. In order to compare the optimal designs, their design efficiencies are compared.

Session 1: 1:10 pm

Current Health Trends and Risk Behavior Analysis in American Youth

Deanna (DeDe) Schreiber-Gregory

The current study looks at recent health trends and behavior analyses of youth in America. Data used in this analysis was provided by the Center for Disease Control and Prevention and gathered using the Youth Risk Behavior Surveillance System (YRBSS). This study outlines demographic differences in risk behaviors, health issues, and reported mental states. Interactions between risk behaviors and reported mental states were also analyzed. Results included reporting differences between the years of 1991 and 2009. All results are discussed in relation to current youth health trend issues.

Break 1 (1:30 - 2:00 pm)

Session 2: 2:00 pm

Using Entropy as a Criterion for Variable Reduction in Cluster Data

Christopher Olson

Entropy is a measure of the randomness of a system state. This quantity gives us a measure of uncertainty that is associated with each particular observation belonging to a specific cluster. We examine this property and its use in analyzing high dimensionality datasets. Entropy proves most interesting in identifying possible dimensions that do not contribute meaningful classification to the clusters present. We can remove the dimension(s) found which are the least important and generalize this idea to a procedure. This algorithm follows the familiar backward or forward selection process that is found in regression literature. After identifying all the dimensions that should be eliminated from the dataset, we then compare its ability in recovering the true classification of the observations versus the calculated classification of the data. From the results obtained, it is clear that entropy is a good candidate for a criterion in variable reduction. The code developed while under this investigation will be rolled into an R package or be made publicly available upon request.

Session 2: 2:20 pm
A Modified Approach to K-Means Using Mahalanobis Distances
Josh Nelson

A problem that arises quite frequently in statistics is that of identifying groups, or clusters, of data within a population or sample. The most widely used procedure to assign data to a set of observations is known as k-means. The k-means algorithm accomplishes this task by first initializing a pre-specified number of cluster centers, usually at random. Then, the points are meted out into clusters by assigning them to the cluster whose center is the closest. Once this step is completed, new cluster means are calculated, and the process begins again. When the clusters no longer change from the one step to the next, the algorithm is said to converge. The main limitation of this algorithm is that it uses the Euclidean distance metric to assign points to clusters. Hence, this algorithm operates well only if the covariance structures of the clusters are nearly spherical in nature. To remedy this shortfall in the k-means algorithm, a new initialization method was introduced, and the Mahalanobis distance metric was used to capture the variance structure of the clusters. If this method serves as a significant improvement over the k-means algorithm, then it will provide a useful tool for analyzing clusters.

Session 2: 2:40 pm
Communications from the Front Lines: Problems from Statistical Consulting
Dr. John Reber (keynote speaker)

Statistical consulting often involves solving nonstandard problems in sub-optimal ways. The problems are often interesting and practical, but due to time constraints a full theoretical treatment is sometimes set aside in favor of a quicker approximate solution. While the client is ideally satisfied with the approximate result, the theoretical questions that are raised can still be of interest to statistical researchers. We'll discuss three recent problems (two from genetics, one from paleontology), along with an approximate solution and potential theoretical work for each of them.

Break 2 (3:20 - 3:50 pm)

Session 3: 3:50 pm

Does Football Momentum Translate into Points?

Michael Price

The average number of points per drive following turnovers is compared to the expected number of points. It is found that the average number of points per drive following turnovers is significantly higher than the expected number of points except when the offensive unit takes over the ball in the red zone after a turnover.

Session 3: 4:10 pm

Examining factors that affect point spread in NBA basketball games

Scot Jones

A stratified random sample of 180 NBA basketball games was taken over a three year period. A model was developed to predict point spread based on various factors. The year the games was played did not matter. The model developed was used to predict whether or not a team won based on knowing the values of various factors for a stratified random sample of 120 games for the 2011-2012 season. The model had an accuracy of 91%.

Session 3: 4:30 pm

Nonparametric Test for the Umbrella Alternative in the Randomized Complete Block and Balanced Incomplete Block Mixed Design

Michael Hemmer

Nonparametric tests have served as robust alternatives to traditional statistical tests with rigid underlying assumptions. The most common motivation for conducting a statistical test is to detect a difference in treatment means. If one expects the treatment means to follow an umbrella alternative, then the test developed in this research will be applicable in the Balanced Incomplete Block Design (Hemmer's test). It is hypothesized that Hemmer's test will prove to be more powerful than the Durbin test when the treatment means follow an umbrella alternative. The Durbin test tests the general alternative in a BIBD. A mixed design consisting of a Balanced Incomplete Block Design and a Randomized Complete Block Design will also be considered. Two test statistics are developed for the umbrella alternative in this design. Monte Carlo simulation studies were conducted using SAS to estimate powers. The first study compared Hemmer's and Durbin's tests. The second study compared the two tests developed for the mixed design. For all scenarios, various underlying distributions were used with 3, 4, and 5 treatments, and a variety of peaks. For the mixed design, cases were considered when the number of complete blocks was equal to, greater than, and less than the number of incomplete blocks. Recommendations are given.

POSTERS

Markovian Approximation and Optimum Statistical Inference of Land-Cover Change at a Transient Watershed: Ramifications on Vulnerability of Arboreal Ecosystems

B. D. Madurapperuma[1], P. G. Oduor[2], L. Kotchman[3]
[1]*Environmental and Conservation Science Program, North Dakota State University, USA*
[2]*Department of Geosciences*, *North Dakota State University*, *USA*
[3]*State Forester*, *North Dakota Forest Service*, *USA*

Forest cover change prediction is crucial for forest managers and for urban developers to make better management of forest resources especially in a highly vulnerable watershed such as Devil's Lake Basin Watershed. This paper presents the dynamics of forest to non-forest (FNF) and vice versa transition using stochastic models and affiliated statistical analyses. National Agricultural Statistic Service (NASS) grid data published by the USDA for years 2001 to 2005 were used to make transition probability matrices for each combination of time steps using SemGrid an open-source software. FNF transition probabilities were further subjected to multivariate analysis. The Detrended Correspondence Analysis (DCA) showed that forest to forest (FF) transition is deviated along the axis 1 of the ordination diagram from other land use class transitions. Three distinguished groups were separated from the ordination diagram: (a) grass, pastureland and grain, hay, seeds (b) idle cropland and row crops (c) water and urban developed. The transition probability for unchanged forest ($.22 \leq Pff \leq 0.67$) was remarkable from 2001 to 2005 for the different time steps. This study offers a glimpse into forests on the edge of a rapidly changing urban and rural environment.

Gender Effects on Child Labor: A Case Study of Malawi and Tanzania

Courage Mudzongo

Child labor is on the increase and this is exacerbating an already desperate situation for children caught up in this trap. The International Labor Organization (ILO) (2010) states that for over a decade, child labor has been recognized as a key human rights concern. Despite the huge social reform movement that has been generated around this issue, an estimated figure of 200 million children worldwide is still trapped. A staggering 115 million at least, are subject to its worst forms and live in Sub Saharan Africa (ILO, 2008; 2010 and UNICEF, 2002). There are many contributing determinates to this crisis such as individual factors and hence the focus on the influence of the child's gender. Using the Chi Square Test, the relationship between gender and child labor has been tested on UNICEF and World Bank datasets. This research is part of a larger project that seeks to uncover the influence of more than just individual factors but contextual factors that lead to child labor. The end goal for this work is to lay a foundation for recommendations of more strategic and relevant interventions for Malawi and Tanzania.