2013 Iowa Chapter Meeting of the ASA

Nov 1, 2013
Ames, IA

Welcome

Welcome to the 2013 IA chapter meeting of the American Statistical Association. We are glad you could join us today and we are delighted about this year's poster presentation. There is a large variety of topics all of which sound exciting. Please talk to presenters about their work and engage in interesting and fruitful discussions.
Most importantly, have fun!

Your Chapter Officers,
Ulrike Genschel, Jeff Jonkman, Joe Lang, Tom Moore

Program

3:45 PM	Registration opens; please pick up your name tag, abstract booklet (and if applicable pay registration fee)
4:00 PM	Welcome
4:15 PM — 5:00PM	Poster Session 1 (Posters # 1 – #13)
5:15 PM — 6:00PM	Poster Session 2 (Posters # 14 – #27)
6:15 PM	Let's celebrate future statisticians and recognize contributing undergraduate and graduate students. (Note prizes will be available!)
6:30PM — 7:00PM	IA Chapter Business Meeting: Election of new Officers
7:00PM	Dinner for everyone who traveled from far away (others are welcome to join)

Participants

Bradley University: David Quigg

Vanward Statistics: Margot Tollefson

Grinnell College: Ana Ovtcharo, Jeff Jonkman, Samantha Mizuni

University of Iowa: Alexandria Bishop, Anna Pritchard, Bo Wang, Colin Lewis-Beck, Fuli Zhang, Jessica Orth, Joseph Lang, Michael Mitsche, Mitch Kinney, Ryne Van Krevelen, Stephanie Kommes, Yingying Liu

Iowa State University: Alicia Carriquiry, Andee Kaplan, Andreea Erciulescu, Andrew Sage, Anna Peterson, Brenna Curley, Bronson Recker, Bryan Stanfill, Dan Nettleton, Dan Nordman, Dave Osthus, Dennis Lock, Divya Mistry, Eduardo Trujillo-Rivera, Emily Casleton, Eric Hare, Geoffrey Thompson, Guillermo Basulto-Elias, Heike Hofmann, Hejian Sang, Hillary Chaney, Jarrod Brockman, Jillian Lyon, Jungyun Yoo, Kenneth Koehler, Kenneth Wakeland, Kevin Kasper, Maggie Johnson, Marie Vendettuoli, Mark Kaiser, Martin Silerio, Matthew Simpson, Matthew Van Hala, Mervyn Marasinghe, Millicent Grant, Philip Dixon, Samuel Benidt, Sarah Nusser, Shu Yang, Stephanie De Graaf, Stephanie Zimmer, Ulrike Genschel, Xiyuan Sun, Yihui Xie, Yaqin Deng, Zhengyuan Zhu

Abstracts

Poster Session 1

1: Using Random Forests to Estimate Win Probability Before Each Play of an NFL Game

Dennis Lock, Dennis.f.lock@gmail.com

Before any play of a National Football League (NFL) game, the probability that a given team will win depends on many situational variables (such as time remaining, yards to go for a first down, field position and current score) as well as the relative quality of the two teams as quantified by the Las Vegas point spread. We use a random forest method to combine pre-play variables to estimate win probability (WP) before any play of an NFL game. When a subset of NFL play-by-play data for the 12 seasons from 2001 to 2012 is used as a training dataset, our method provides WP estimates that resemble true win probability and accurately predict game outcomes, especially in the later stages of games. In addition to being intrinsically interesting in real time observers of an NFL football game, our WP estimates can provide useful evaluations of plays and play calls.

2: An Introduction to Dynamic Documents and knitr

Yihui Xie, xie@iastate.edu

Traditionally we finish statistical computing before writing reports, and the results written into the reports are essentially "dead." To update the reports, we have to redo computing. In this talk, we introduce live documents that generate reports directly, with results dynamically obtained from computer code. We show how we can change the research workflow and present results in new ways with the help of tools such as RStudio + knitr and modern technologies like HTML5 and WebSockets. Reproducible research should be so natural that we do not even notice it.

3: Determining Tightest Cell Bounds in Rounded Contingency Tables

Andrew Sage, ajsage@iastate.edu

Displaying cell counts or exact frequencies in a contingency table presents a disclosure risk, potentially violating the privacy of subjects under study. When rounded frequencies are displayed, this risk is sometimes alleviated. We present an approach that uses dynamic programming to quickly determine the tightest bounds on individual cell counts given only rounded conditional frequencies and sample size. This algorithm can be used by data providers to determine the disclosure risk of tables released to publicly.

4: Putting Down Roots: A Graphical Exploration of Community Attachment

Andee Kaplan, ajkaplan@iastate.edu

In this presentation we explore the relationships that individuals have with their communities. We present our findings using interactive as well as static visualizations. This work was prepared as part of the ASA Data Expo 2013 sponsored by the Graphics Section and the Computing Section, using data provided by the Knight Foundation Soul of the Community survey. The Knight Foundation in cooperation with Gallup surveyed 43,000 people over three years in 26 communities across the United States with the intention of understanding the association between community attributes and the degree of attachment people feel towards their community.

 # 5: Can you buy a president? Politics after the Tillman Act

Eric Hare, erichare@iastate.edu

Motivated by the 2010 Citizens United ruling and the subsequent birth of "Super PACs", this paper uses independent expenditures data from the Federal Elections Commission, in conjunction with presidential polling data to analyze the 2012 presidential campaign. Using R, and several packages, we scrape data from these sources and analyze them in order to highlight interesting trends in campaign spending. Furthermore, we correlate these trends in spending over time to the changes in the polls. Ultimately, there is not a lot of evidence to support a clear and direct relationship between increases in spending and changes in public support. However, our analysis does reinforce some commonly held views of Super PAC spending habits and the candidates' geographical areas of strength and weakness.

 # 6: Improved Confidence Regions for the Central Orientation in $SO(3)$

Bryan Stanfill, stanfill@iastate.edu

Data as three-dimensional rotations have applications in computer science, kinematics and materials sciences, among other areas. Inference for the central orientation (mean) from a sample of such data is an important problem and has received increased attention in the literature, e.g. Rancourt et al. (2000) and Bingham et al. (2009). Currently, much of that attention has come from a parametric standpoint, which is only valid for large samples from a class of distributions that behaves normally in $SO(3)$. In this poster we offer non-parametric bootstrap procedures which achieve coverage rates closer to the nominal level under more general conditions. We also clarify and extend an asymptotic result that motivates the parametric intervals already in the literature. Our methods are illustrated alongside our competitors' in a simulation study and data example.

 # 7: Model averaging predictions is good; model averaging multiple regression coefficients is bad.

Philip Dixon, pdixon@iastate.edu

Model averaging, or multi-model inference, avoids the false sense of precision arising when inferences are conditional on a model chosen from data. It accounts for the uncertainty due to model choice. It is widely used in mark-recapture studies to estimate a population size, N, while accounting for uncertainty in the capture probability model. It has since been widely used with model selection in multiple regression. Instead of reporting coefficient estimates for the "best" model, estimates are averaged over a collection of models. I argue that there is a fundamental difference between these two uses of model averaging. In mark-recapture, N is consistently defined in all models. In multiple regression, the interpretation of a specific regression parameter changes between models because the suite of "variables held constant" changes. Using a variety of examples, I show that conditioning on the best model leads to better parameter estimates than model averaging in the multiple regression setting. If the focus is on predicted values, which are consistently defined across models, instead of regression coefficients, model averaging does lead to better predictions.

 # 8: Exploring Distributions of Seasonal Climate Forecasts

Kenneth Wakeland, wakeland@iastate.edu

The skill of long-range weather forecasts, from one to five months into the future, is typically assessed using values aggregated (averaged) over large expanses of space such as a state or region and moderate periods of time such as a month or season. But the process of producing such forecasts produces many individual

values in space and time, values that can be used to define a variety of empirical distributions. We present a number of statistical methods that can be used to investigate the characteristics of such distributions, ranging from simple histograms in space and time to complex hierarchical statistical models that incorporate the effects of individual forecast runs, latitudinal gradients, and temporal separation between model run and forecast target. We illustrate some of these methods with 30 years of forecasted daily maximum July temperatures in Iowa, each year of which has forecast runs started at 4 times in each of 5 days for each of the months from February through June.

 # 9: Shattering sub-permutations in an array of n-permutations
Stephanie De Graaf, sdegraaf@iastate.edu

We consider an array of k permutations on $[n]$ where each row is a permutation, and we look for permutation patterns present in subsets of columns. We wish to identify k, the minimum number of rows required to obtain all order-isomorphic t-permutationsin the rows of any choice of t columns. Using a probabilistic method with the Lovasz Local Lemma, we improve the known upper bound for k. We also obtain an upper bound when requiring multiple copies of each permutation pattern to be present in each set of t columns.

 # 10: A Bootstrap Confidence Interval for Meta-Analysis
Ana Ovtcharo and Samantha Mizuni, ovtcharo@grinnell.edu, mizunosa@grinnell.edu

Meta-analysis refers to the methodology used to analyze multiple independent but related studies. By pooling the studies it is possible increase the statistical power of the meta-analysis beyond that of the individual studies so that the overall effect might be detected. Like all forms of statistical analysis meta-analysis relies upon several assumptions. These are not always met, potentially creating problems with the overall analysis and conclusion. Non-parametric methods, such as the bootstrap, do not rely upon assumptions of normality, and thus are more robust under non-normal conditions than parametric methods. We propose a bootstrap method for creating confidence intervals for random-effects meta-analysis. The approach is very sensitive to the size of the meta-analysis although not to the value the amount of heterogeneity between studies. Under normal circumstances, it has a higher error rate than either of two normal methods it was compared against, the DerSimonian and Laird approach or the Sidik and Jonkman method.

 # 11: Electron Flux, Solar Wind Speed, Sunspots, and Dynamic Linear Models
Dave Osthus, dosthus@iastate.edu

In the field of space weather, forecasting relativistic electron flux (electron flux) in the radiation belt is a difficult problem, in large part because the dynamics governing space weather are not well understood. Much effort has been invested in both trying to understand the dynamics governing space weather as well as forecasting electron flux. The former informs the latter as more is learned about the relationships between electron flux and leading covariates (e.g. solar wind speed), those relationships can be built into a forecasting model. We, however, propose an approach where the latter can help inform the former. Through the use of dynamic linear models with time-varying, contextually meaningful parameters, there is the ability to examine how relationships between electron flux and other covariates evolve over time. We illustrate this approach by modeling the relationship between electron flux and solar wind speed, drawing connections with phases of the solar cycle.

 # 12: Modeling Bean Pod Mottle Virus Using Binary Markov Random Fields With Absorbing States

Mark Kaiser, mskaiser@iastate.edu

In a sequence of binary random fields over time, absorbing states occur when a positive value cannot revert to s zero value at later points in time. This happens with plant disease in soybean fields. Absorbing states violate a condition that is used in the construction of Markov random fields. We show how to circumvent this problem and model the spread of bean pod mottle virus over time in an agricultural field.

 # 13: An Analysis of World Life Expectancy

Hillary Chaney and Bronson Recker, htchaney@iastate.edu, brontego@gmail.com

People today live much longer than people did 100 years ago and new breakthroughs in medical technology and understanding of the food and beverages we consume result in an increased life expectancy. In the following we explore how different aspects of our daily lives might potentially affect the average life expectancy around the world. The areas we study include tobacco use, alcohol consumption, population growth, and government spending on health care. We have data available from 1990 to 2006 which were collected by the World Health Organization and which were converted into documents that could be run through R by the website visualizing.org.

Poster Session 2

 # 14: Empirical Likelihood for Irregularly Located Spatial Data

Matthew Van Hala, mvanhala@iastate.edu

Empirical likelihood formulates a likelihood nonparametrically, but has properties analogous to parametric likelihood, such as chi-squared limits for likelihood ratio statistics. Empirical likelihood has previously been extended to time series and spatial lattice data using data blocking techniques. We develop a block-based empirical likelihood method for irregularly located spatial data, present distributional results for the empirical likelihood ratio, examine the performance of the method by simulation, and apply the method to a real data example.

 # 15: Tornadoes in Arkansas: A Network Analysis Approach

Emily Casleton, casleton@iastate.edu

Network analysis is a fast-growing area of research due to the variety of applications that can be modeled as a network. A set of nodes and their relations defines a network. The nodes and relations, i.e., edges, can symbolize a wide range of objects and a network can represent complex patterns of connections and dependencies between them. A new model, the Local Structure Graph Model (LSGM), has been developed to quantify and compare important and interesting features of networks. The main components of the LSGM are a conditional specification and an explicit definition of neighborhoods. Two sets of parameters control the model; one represents the large-scale structure of the network while the other represents the small-scale structure. The parameters and other characteristics of the model will be demonstrated on a dataset of cited tornadoes in Arkansas during April, 2011. The goal of the analysis is to illustrate how the model is able to capture the local structure of the network.

 # 16: Dynamic Graphics: An Interactive Analysis Of What Attaches People To Their Communities

Jessica Orth, jessica-orth@uiowa.edu

In this research, we will investigate several different approaches and methods to displaying multivariate data. Emphasis will be placed on end-user-customization tools and flexibility in dynamic and interactive displays. Specifically, we will highlight the use of motion charts using Markus Gesmann's googleVis package in R. We will demonstrate the visualization of time-series data and also the results of Multidimensional Scaling and Principal Component Analysis using this tool. The goals of these displays are ease of usability and interpretation, dynamic customization options, and the ability to display multivariate data in a meaningful way. We will use data collected from the Knight Foundation and Gallup during the years 2008-2010 to illustrate the attachment of people to their communities in a new and innovative way.

 # 17: Association Between Intake of Added Sugars with Nutrient Intakes for Children and Adolescents Ages 9-18 years old

Brenna Curley, curleyb@iastate.edu

The term "empty calories" refers to the calories that are contributed by some starchy foods, saturated fats, alcohol, and refined sugars (Jenkins, 2004). The effect of empty calories on a person's diet is at odds with maintaining a healthy lifestyle. Data for individuals aged 9 to 18 years old in the United States from NHANES 2003-2008 were used to estimate the association between intake of added sugars and discretionary fats with intake of essential nutrients. We fit a regression model that allows for non-independent measurement error between the dependent and the response variables, to account for the fact that observed daily intakes are noisy measurements of usual intakes. The response variable in our models is the nutrient density (units of the nutrient per 100 calories); calories from empty calories are similarly scaled for consistency. Other covariates in the model (e.g., BMI and age) are assumed to be measured with no error. For certain age-sex groups, added sugars are found to be negatively associated with intake of some nutrients, suggesting that intake of foods with high content of added sugar displace consumption of some nutrients.

 # 18: Improved Interval Estimation of a Comparative Treatment Effect

Ryne Van Krevelen, ryne@vankrevelen.com

Comparative experiments are ubiquitous in scientific research, and there is no shortage of papers that show how to estimate a treatment effect. The object of inference, which is a measure of the treatment effect, is dictated by what assumptions the researcher is willing to make. In many cases neither the object of inference nor the assumptions are clearly presented. This can cause confusion about the scope of the inference. This poster demonstrates one possible method for constructing confidence intervals for treatment effects. This method lends itself to clear descriptions of the assumptions and the object of inference and can be widely applied. Using data from a real world experiment, this method is compared to a Fisher-type randomization interval and to a t interval. Results of a small-scale simulation study will also be presented.

 # 19: Bivariate kernel deconvonvolution approach to estimate the joint density of noisy random variables with unknown error distribution

Guillermo Basulto, basulto@iastate.edu

Replicate observations of 25(OH)D (biomarker for vitamin D status) and iPTH are available on a sample of individuals. We assume that measurements are subject to non-normal measurement error. We estimate

the joint density of these bivariate data via nonparametric deconvolution. The estimated density is used to compute statistics of public health interest, such as the proportion of persons in a group with 25(OH)D values below iPTH, or the value of 25(OH)D above which iPTH is approximately constant. We use a bootstrap approach to compute confidence intervals. Several bivariate kernel density estimators for the noisy data and estimators for the characteristic function of the error are compared.

 # 20: Evaluating the Impact of Nonsampling Errors on Erosion Estimates for the Conservation Effects Assessment Project

Andreea Erciulescu, andreeae@iastate.edu

The Conservation Effects Assessment Project (CEAP) is a series of surveys intended to evaluate environmental outcomes associated with conservation practices. Erosion is one such environmental outcome of interest, because it affects soil quality on cropland and crop productivity, water quality and quantity, and air quality. Errors produce biased erosion estimators at different geographical levels, identified by Hydrologic Unit Codes (HUCs). We discuss possible ways to evaluate effects of two sources of nonsampling error on erosion estimates for the eight digit HUCs.

 # 21: Graphical Lasso Applications for Gene Set Analysis

Kevin Kasper and Mervyn Marasinghe, kmkasper@iastate.edu and mervyn@iastate.edu

The graphical lasso is a method that estimates a sparse covariance matrix and thus can be used to estimate a graph structure. Using the graphical lasso on a multivariate gene set the conditional dependencies that may exist among genes in a set can be estimated. A test for differential expression is proposed for gene set studies with two experimental conditions using the covariance matrix estimated from the graphical lasso. Performance of this procedure is then compared to previously published methods.

 # 22: Modeling Spatial Binary Fields over Time with Dynamic Markov Random Fields

Kenneth Wakeland, wakeland@iastate.edu

Any number of problems in ecology and the environmental sciences, such as monitoring the presence/absence of a species, involve the observation of spatial binary random fields at a sequence of points in time. There is often insufficient information about the scientific processes involved to incorporate a deterministic component for time evolution into a model. We consider Markov random field models with binary conditional distributions that include a stochastic evolution over time based on autoregressive structure for the large-scale model component. These models retain the flexibility of static Markov random field models for representation of spatial dependence in the small-scale model component. Bayesian estimation is accomplished through the use of what has been called the 'double Metropolis algorithm', which requires generation of auxiliary random fields, but does not require the use of perfect sampling. Use of the model is illustrated with simulated and real data.

 # 23: SimSeq: Data Based Simulation of RNA-Seq Data

Samuel Bendit, sbenidt@iastate.edu

RNA-Seq analysis methods are often derived by relying on hypothetical parametric models for read counts that are not likely to be precisely satisfied in practice. Methods are often tested by analyzing data that have been simulated according to the assumed model. This testing strategy can result in an overly optimistic view of the performance of an RNA-Seq analysis method. Rather than generating data from a

questionable parametric model, the SimSeq package provides tools for simulating RNA-Seq data by subsampling from existing datasets. The vector of read counts simulated for a given experimental unit has a joint distribution that closely matches the distribution of actual RNA-Seq data. Users can control the proportion of genes simulated to be differentially expressed (DE) and can adjust the magnitude of differences for DE genes. SimSeq requires a matrix of RNA-Seq read counts with large sample sizes in at least two treatment groups. Although a full version of such a dataset is not distributed with the package, numerous candidate datasets are publicly available.

24: Senior Projects in Statistics

David Quigg, quigg@fsmail.bradley.edu

Since I am frequently the faculty mentor for a students senior project, I find that my own understanding is also inspired by the final result. Specifics are provided for three recent senior projects:

- "Math on the Radio Zipfs Law" with E. Bartosh
- "Card Counting in Blackjack" with N.Dell
- "Logistic Regression: The License Plate Problem" with M.Meyer

25: Modeling Crash Frequency Data

Eduardo Trujillo-Rivera, eduardo@iastate.edu

We propose two Bayesian models based on a small domain approach to describe crash frequency data from Iowa roads. It is of our interest to study the impact of different Covariates on the frequency of crashes and to estimate Safety Performance Functions.

26: Statistical Models for Global Phenological Phenomena

Maggie Johnson, majohnso@iastate.edu

Understanding the effects of phenological events, due to both natural and man-made causes, is critical for research in global climate modeling and agriculture, among many others. During the last two decades, remote sensing data on satellite derived biophysical variables (such as chlorophyll content) have become widely available through the launch of satellites such as MERIS (MEdium Resolution Imaging Spectrometer). Weekly MTCI aggregates from 2003 to 2007 were used to model phenological changes in southern India. Three modeling techniques – a classical time series model with seasonality represented by using Fourier terms, a method integrating the decomposition of time series into season, trend, and white noise components with methods for detecting significant changes (BFAST), and a hierarchical model incorporating spatially distributed covariates such as land use and elevation – were used to extract the phenological variables of onset of greenness, peak of greenness, and end of senescence using an iterative search. The advantages and shortcomings of the three methods are compared and discussed in terms of phenological variable extraction and efficacy across spatial locations.

27: The Mismatch between Existing Statistical Methodology and Classroom Data

Jillian Lyon, jdlyon@iastate.edu

As research in education becomes more prominent, many different pedagogical approaches and classroom techniques have been adopted and explored for their effectiveness. However, adequately demonstrating the

efficacy of such instructional innovations by adhering to necessary statistical rigor has proven to be difficult due to the nature of data from educational settings. For example, students' performance data obtained from assessments can generally not assumed to be independent, especially when collected from students in the same classroom. Still a common misconception among researchers, especially outside the field of statistics, is to treat students as the experimental unit instead of the classroom. Additionally, published research in a variety of fields often fails to address potential instructor effects that are often inevitably confounded with treatment. We discuss these difficulties and highlight them through two simulation studies. Our goal is to underscore the need for new statistical methodology to analyze classroom data as educational researchers will continue to face non-ideal data in the form of small sample sizes, lack of replications and dependencies among observations.