Fall 2013 CO/WY ASA meeting
Anschutz Medical Campus, Ed 2 North Building, Room P28-2104
Friday, October 18, 2013
Time: 12:00-3:30pm (lunch provided)

Draft of Agenda:

12:00-12:20pm	**Welcome and lunch (nominations for officers)**
12:20-12:40pm	**Lunch session: Matt Pocernich, Neptune.** *Reproducible Research for Statisticians.*
12:40- 1:00pm	**Bruce Bugbee, CSU.** *Variational approximations applied to semiparametric regression.*
1:00- 1:20pm	**Elizabeth McClellan, MSU.** *NetWeAvers: A Statistical Method for Integrative Biological Network Analysis.*
1:20- 1:40pm	**Miranda Kroehl, UCD.** *An Evaluation of Permutation Methods for Testing Mediation in the Presence of Covariates.*
1:40- 1:50pm	**Break**
1:50- 2:10pm	**Business meeting, elections**
2:10- 2:30pm	**Daniel Yorgov, UCD.** *Combined Association and Admixture Mapping for Complex Traits.*
2:30- 2:50pm	**Karl Ellefsen, USGS.** *Mixture-model clustering of regional geochemical data.*
2:50- 3:10pm	**Xiyue Liao, CSU**. *coneproj: An R Package for the Primal or Dual Cone Projections with Routines for Constrained Regression*.
3:10- 3:30pm	**Annie Lu, Samantha Estrada M, Steven Pulos, UNC**. *Psychometric Evaluation of the Revised Current Statistics Self-efficacy (CSSE-30) in a Graduate Student Population Using Rasch Analysis*.

Abstracts:

Bruce Bugbee, CSU. *Variational approximations applied to semiparametric regression.*

The assumption of a constant variance term is considered standard for a wide range of regression models. However, this assumption can cause incorrect estimation of uncertainty regarding parameter estimates if the data has a non-constant variance structure. Our work focuses on the case of the estimation of flexible semiparametric models in the presence of non-constant errors. Specifically, we focus on providing a fast deterministic approximation as of the model using non-conjugate variational inference.

Elizabeth McClellan, MSU. *NetWeAvers: A Statistical Method for Integrative Biological Network Analysis.*

"The discovery of functionally related groups from a set of significantly abundant proteins or highly expressed genes is an important step in a proteomics or transcriptomics analysis pipeline. In this talk I will describeNetWeAvers (Network Weighted Averages) for analyzing groups of regulated proteins or genes (called "features"), e.g. as defined by clusters of protein-protein interactions. NetWeAvers has been implemented in an R package that provides a novel method for analyzing feature-level data integrated with biological networks. The method includes an algorithm for finding dense clusters of features and a permutation algorithm to calculate cluster *p*-values. Optional steps include summarizing quantified peptide values to single protein values (for proteomics data) and a differential expression test. I will show that NetWeAvers applied to a public proteomics dataset finds statistically significant and biologically relevant clusters."

Miranda Kroehl, UCD. *An Evaluation of Permutation Methods for Testing Mediation in the Presence of Covariates.*

The evaluation of mediating variables is becoming increasingly popular in clinical research. Often, researchers are interested in determining whether there is an indirect effect of the independent variable, X, acting through the mediating variable, M, on outcome, Y. A variety of methods have been proposed to estimate and test the indirect effect, with the two common approaches being an estimate obtained from the product of two regression coefficients, or a test of joint significance of those coefficients. Recently, Taylor and MacKinnon (2012) considered different applications of permutation testing for use in a single-mediator model; however covariates were not included in this evaluation. Permutation methods work by breaking up associations between variables of interest by random shuffling of the data. Often, the raw data can be permuted. However with multiple regression models, it is often more appropriate to permute residuals obtained from fitting a reduced model of the data. Because an important assumption in mediation analysis is one of no unmeasured confounders, we extended the work of Taylor and MacKinnon to allow for the presence of a covariate, C, by permutation of residuals. We consider three scenarios: (1) C is a covariate, associated with outcome Y but not with X or M, (2) C is a confounder of the X-Y relationship, and (3) C is a confounder of the M-Y relationship. We considered a permutation test of the product under a reduced model, a permutation test of joint significance under the reduced model, and a permutation test of the product under the full model. We compare the different methods on Type I error rates and power for each X2 scenario, and make recommendations on mediation analysis using permutation testing when covariates are present.

Daniel Yorgov, UCD. *Combined Association and Admixture Mapping for Complex Traits.*

Combined admixture and association testing in recently admixed populations might elevate the power of detection of complex traits by merging two complimentary sources of genetic signal. I will present a method for joint analysis of ancestral and genotype effects at a single locus for quantitative trait. I will describe the steps necessary to carry out such a study from selection of reference populations and local ancestry inference through the testing itself. A significant locus for diastolic blood pressure was found when applying this method to Mexican American sample of unrelated individuals.

Karl Ellefsen (with D.B. Smith and J.D. Horton), USGS. *Mixture-model clustering of regional geochemical data*.

Mixture-model clustering of regional geochemical data is a statistical procedure that is useful for interpretation. Because geochemical data are a type of compositional data, straightforward application of standard statistical procedures can yield erroneous results. Thus, we have developed and implemented (in the R statistical programming language) a robust clustering procedure that accounts for the compositional properties of the data: All element concentrations are first transformed with the isometric log-ratio transformation. The transformed concentrations are then used to calculate robust principal components. These components are clustered using a mixture model for which the probability density functions are multivariate normal, and the conditional probabilities that a sample is related to the density functions are calculated. In addition, random samples are drawn from each of the density functions and then are back-transformed to equivalent element concentrations.

The clustering procedure is evaluated with soil geochemical data from a survey of the state of Colorado (United States of America). The data comprise 959 samples with 31 element concentrations for each sample. The chosen mixture model has 4 density functions, and the calculated conditional probabilities partition the 959 samples into 4 clusters. For each cluster, most samples are spatially close together and thus are related to specific geologic features such as surficial deposits or bedrock. The independently-known geochemical properties of these geologic features are consistent with the random sample concentrations, and the order statistics for the random sample concentrations are almost identical to the corresponding order statistics for the field data (i.e., the measured concentrations for those samples with high conditional probabilities). Both results suggest that the clustering procedure is accurate. Another benefit of mixture-model clustering is that the element concentrations for each cluster are approximately statistically stationary, making them suitable for additional statistical processing such as multivariate kriging.

Annie Lu, Samantha Estrada M, Steven Pulos, UNC. *Psychometric Evaluation of the Revised Current Statistics Self-efficacy (CSSE-30) in a Graduate Student Population Using Rasch Analysis*.

Statistics serves as a powerful research tool nowadays in many scientific fields. Various statistical education studies have been done on the undergraduate student population, but not extensively on the graduate student population. However, statistical training is equally important for graduate students given that the majority of them will become future researchers. Self-efficacy has been proposed as a strong predictor of students' academic performance. The current statistics self-efficacy (CSSE) scale, developed by Finney and Schraw (2003), is a 14-item instrument to assess student's statistics self-efficacy. No previous research has used Rasch modeling to evaluate the psychometric structure of its scores at the item level, and only a few of them have applied the CSSE in a graduate school setting.

A modified 30-item CSSE scale was tested on a graduate student population (*N*=136). The Rasch rating scale analysis identified 26 items forming a unidimensional measure. Assumptions of sample-free and test-free measurement were confirmed, showing scores from the CSSE-26 are reliable and valid to assess graduate students' level of statistics self-efficacy. The finding is important for statistical education given that the CSSE-26 could serve as a useful tool for those studies examining the effect of statistics self-efficacy on statistics performance. Further, professors could utilize the CSSE-26 to (1) identify graduate students with lower statistics self-efficacy, (2) identify statistical concepts that relate to their low self-efficacy, (3) provide these students help or interventions to build their confidence, and (4) explore alternative teaching techniques to explain certain statistical concepts. Proper statistical training of graduate students will have a positive impact on the research community as a whole.

Xiyue Liao, CSU. *coneproj: An R Package for the Primal or Dual Cone Projections with Routines for Constrained Regression*.

The **coneproj** package contains routines for cone projection and quadratic programming, plus applications in estimation and inference for shape-restricted regression. For the **coneA** and **coneB** functions, the vector to project is provided by the user, along with the cone specification and a weight vector. For **coneA**, a constraint matrix is specified to define the cone, and for **coneB**, the cone edges are provided. The **coneA** and **coneB** algorithms have been coded and compiled in **C++**, and are called by **R**. The **qprog** function transforms a quadratic programming problem into a cone projection and calls **coneA**. The **constreg** function does estimation and inference for parametric least-squares regression with constraints on the parameters (using **coneA**). A p-value for the "one-sided" test is provided. The **shapereg** function uses **coneB** to provide a least-squares estimator for a regression function with several choices of constraints including isotonic and convex regression functions, as well as estimates of parametrically modeled covariate effects. Results from hypothesis tests for significance of the effects are also provided. This package is now available from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=coneproj.

Matt Pocernich, Neptune. *Reproducible Research for Statisticians*.

As a concept, nearly everyone would agree that science should be reproducible. In practice, things are less clear cut and much more gray. Drawing on experiences in science, engineering and programming, this talk presents a range of definitions for reproducible research and discusses a range of practices. For consumers of reports, studies and analyses - it is nice to be able to rely on more than faith that an analysis is correct. As producers, when your work is scrutinized (i.e. when the shit hits the fan) - it is nice to be able to clearly illustrate how results were produced. This talk briefly discusses the concept of reproducibility in different fields and some ideas and practices we statisticians can borrow from the field of software engineering and quality control to improve our work.