2016 Seminars

12/07/2016

The Evolution of a Statistical Consulting Course

Colleen Kelly, PhD

Abstract: In 2000, Duane Steffey and I founded the SDSU Consulting Center and developed a Statistical Consulting course in response to university and private consulting requests and a desire to better train our graduate students for careers as applied statisticians. In the successive 15 years, I became increasingly devoted to statistical consulting as a career and eventually left academia for a career in consulting. In 2009, I founded Kelly Statistical Consulting. My industrial consulting experience has revised my vision of what is important to teach in a consulting course. In this talk, I present the common elements to most statistical consulting courses and how my presentation of these elements has evolved over the last 15 years. I discuss the (sometimes hard) lessons learned, and what I believe to be the key elements of a successful course.

Bio: Dr. Colleen Kelly is an Accredited Professional Statistician™ and has over 25 years of statistical consulting experience as a statistical consultant, professor, and researcher specializing in statistical methodology for clinical trials and biotechnology. As a tenured associate professor of statistics at San Diego State University, Dr. Kelly co-founded and co-directed the university’s statistical consulting center. At Victoria University in Wellington, New Zealand, she directed the university’s statistical consulting center and developed and taught their statistical consulting course. Currently, she heads Kelly Statistical Consulting, Inc., which provides statistical services to biotechnology, pharmaceutical and medical device companies.

11/02/2016

A distance-weighted model for methylation change with application to whole genome bisulfite sequencing data

Michelle Lacey, PhD

Abstract: Variation in cytosine methylation at CpG dinucleotides is often observed in genomic regions, and analysis typically focuses on estimating the proportion of methylated sites observed in a given region and comparing these levels across samples to determine association with conditions of interest. While sites are typically treated as independent, when observed at the level of individual molecules methylation patterns exhibit strong evidence of local spatial dependence. We previously introduced a neighboring sites model to account for correlation and clustering behavior observed in two tandem repeat regions in a collection of ovarian carcinomas. We now introduce an extension of the model that accounts for the effect of distance between sites. We apply our model to data from a whole genome sequencing experiment using overlapping 300-bp reads, demonstrating its ability to detect distance-weighted effects in regions with intermediate levels of methylation.

Bio: Michelle Lacey earned her PhD in Statistics from Yale University and joined the Tulane faculty in 2003. She is currently appointed as Associate Professor of Mathematics and Adjunct Associate Professor of Biostatistics at Tulane University, and in addition to regularly teaching graduate courses in statistical modeling and data analysis for the School of Science and Engineering she is a contributing lecturer for courses at the Tulane University School of Medicine and the School of Public Health and Tropical Medicine. Dr. Lacey directs the Tulane Cancer Center Genomics Analysis Core to provide statistical support to researchers conducting high-throughput experiments, and she maintains an independent research program in epigenetic modeling and analysis. She also collaborates with researchers in the school of Science and Engineering and has recently established a consulting relationship with the World Food Programme to assist in the development of statistical methods for modeling and analysis of food security survey data.

10/26/2016

A flexible framework for fitting mixed models based on automatic differentiation

Hans J. Skaug, PhD

Abstract: I will describe a flexible framework for doing empirical Bayes inference in general mixed models. The marginal likelihood is evaluated using the Laplace approximated, and optimized using a Newton-type method. The technical details of the Laplace approximation is hidden from the user via a technique called Automatic Differentiation. The approach has been implemented in the software package TMB (https://github.com/kaskr/adcomp). TMB is an R package, but links to C++ code for evaluation of the joint (in fixed and random effects) likelihood. I will discuss how TMB can be used to build mixed model R packages, and give examples of such R packages.

Bio: Hans J. Skaug is Professor in statistics at the Department of Mathematics, University of Bergen. He received his PhD (Dr. Scient) in 1994. His field of research is statistical ecology and computational statistics.

10/05/2016

Equivariant Functional Analysis of Curves in SO(3) with Applications to Gait Analysis

Fabian Telschow, PhD

Abstract: In gait analysis of the knee joint data are curves in the group of 3×3 rotation matrices. We introduce and study S-equivariant functional models (viz., Gaussian perturbations of a center curve) and provide a uniform strongly consistent estimator for the center curve. Here S is a certain Lie group, which models the effect of different marker placements and self-chosen walking speeds in real gait data. Moreover, we provide novel estimators correcting for different marker placements and walking speeds and provide different statistical tools to analyze such data, for example, simultaneous confidence sets and permutation tests. The methods are applied to real gait data from an experiment studying the effect of short kneeling.

Bio: Fabian Telschow got his PhD degree from the University of Goettingen. His supervisor was Stephan Huckemann. He developed statistical tools for the analysis of biomechanical gait data in cooperation with the biomechanist Michael Pierrynowski from McMaster University, Canada. During his studies for the Msc. degree in pure math in Göttingen (specialized in the intersection of algebraic topology and geometry) he worked part-time in the group of Axel Munk as a student research assistant on statistical analysis of 2D-NMR spectroscopy. His current research interests are real world applications of non-euclidean statistics, especially, if the data are curves.

09/28/2016

A statistical perspective on health behavior research: a tale of multiple exposures and multiple outcomes

Loki Natarajan, PhD

Abstract: Leading a healthy lifestyle can positively impact health, and reduce the risk of cancer, cardiovascular disease, and other chronic diseases. Health behaviors include many modifiable factors such as physical activity, diet, sleep, and smoking. This multiple exposure-multiple outcomes aspect of health behavior research calls for novel statistical approaches for study design and data analysis. In this talk we will discuss some of these approaches.

In the first part of the talk, we will present a “biobridge” design for a lifestyle intervention trial. Specifically, we will develop an analytic method to calculate a weighted risk score from several intermediate outcomes, and discuss how to quantify future clinical benefit through intervention-related changes on this risk score. Relative weights for the intermediate outcomes are derived by comparing a disease model conditional on the joint distribution of these outcomes to the corresponding marginal models. We will show analytically and via simulations that using marginal parameters as the weights in the risk score, and ignoring inter-correlations amongst the outcomes, yields biased estimates. Our proposed weighted risk score corrects for these biases. We will apply this method to design a weight-loss intervention trial with multiple biomarker outcomes.

In the second part of the talk, we will discuss the use of Bayesian networks to model multiple health behaviors and outcomes. Bayesian networks are a probabilistic machine learning approach which can be used to model multivariate relationships and represent them via intuitively meaningful graphs. We will apply this method to a sample of 333 overweight post-menopausal breast cancer survivors to model associations between BMI, lifestyle behaviors (alcohol intake, smoking, physical activity, sedentary behavior, sleep quality), psychosocial factors (depression, quality of life), biomarkers (insulin, C-reactive protein), demographics (age, education), and tumor factors. Using these networks, we will quantify the strength of association and infer (conditional) dependencies amongst these variables. Our results demonstrate that Bayesian networks could be a powerful exploratory tool for health behavior research.

09/07/2016

Exact inference on the restricted mean survival time

Lu Tian, ScD

Abstract: In a randomized clinical trial with the time to event as the primary endpoint, one often evaluates the treatment effect by comparing the survival distributions from two groups. This can be achieved by for example estimating the hazard ratio under the popular proportional hazards (PH) model. However, when the hazard rate is very low, e.g., in safety studies, there may be too few observed events to warrantee the valid asymptotical inferences under the PH regression. The exact inference including hypothesis testing and constructing 95% confidence interval for the treatment effect is desired. In this paper, we have developed exact inference procedure for estimating the treatment effect based on the difference in restricted mean survival time between two arms, which is more appealing than hazard ratio in many applications. The proposed procedure is valid regardless of the number of events. We have also performed a simulation study to examine the finite sample performance of the proposed method.

Bio: Dr. Lu Tian received my Sc.D. in Biostatistics from Harvard University. He has considerable experience in statistical methodological research, planning large epidemiological studies, performing data management for randomized clinical trials and conducting applied data analysis. My current research interests are in developing statistical methods in personalized medicine, survival analysis, meta analysis and high throughput data analysis.

07/06/2016

How Principles for Analyzing Incomplete Data Motivate Viewing Trust and Understanding as Twin Pillars of Ethics in Statistics

Thomas R. Belin, PhD

Abstract: Accepting the need to enunciate ethical principles in the field of statistics, how might it be possible to encompass the scope and generality of what we do into a complete yet digestible set of guidelines? Drawing on reflections by leading statisticians about the nature of our work, scientific insights regarding how the human condition induces imperatives for people to communicate with one another, game-theory perspectives on competition and cooperation, and other philosophical discourse on the ethics of interpersonal interactions, it is argued that trust and understanding are essential core principles that can serve as the basis for judging whether a statistical approach is ethical. The framework's simplicity makes it easy to communicate, its generality gives it power, and its positive-sum appeal could be used to promote professional identity development around ethics. The presentation will also consider connections between this framework and principles for analyzing incomplete data, where the dual goals of reflecting all available information and accurately representing uncertainty have parallels to cultivating understanding and cultivating trust. Recent efforts to develop flexible joint-modeling strategies to handle highly multivariate data sets with a broad array of data types will also be discussed.

Bio: Thomas R. Belin, Ph.D. is a Professor in the UCLA Department of Biostatistics with a joint appointment in the UCLA Department of Psychiatry and Biobehavioral Sciences. He started at UCLA in 1991 after receiving his Ph.D. that year from Harvard University, working with Donald Rubin in the Harvard Department of Statistics on incomplete-data problems related to the decennial census in the United States. Specializing in statistical analysis with missing data and related extensions to causal inference, he has supervised over a dozen doctoral dissertations and was recognized in 2015 by the UCLA Public Health Student Association for "Outstanding Advising and Mentorship for Ph.D. and Dr.P.H. Students". He also serves as Vice Chair of the UCLA Department of Biostatistics, and his professional activities include being a member since 2014 of the American Statistical Association Committee on Professional Ethics. He was elected Fellow of the American Statistical Association in 2004, and in 2005 he received the Washington (D.C.) Statistical Society Gertrude M. Cox Award honoring a statistician making "significant contributions to statistical practice."

05/18/2016

Statistical Investigation of Ensemble Kalman Filter

Soojin Roh, PhD

Abstract: Data assimilation is a statistical method to combine the output from numerical models with observations to give an improved forecast. The ensemble Kalman filter is a widely used data assimilation method in diverse areas such as weather forecasting and aerospace tracking. In this talk I will discuss the ensemble Kalman filter and some practical issues. I will then discuss a robust ensemble Kalman filter.

Bio: Dr. Soojin Roh received her PhD in Statistics from Texas A&M University. She is currently a lecturer in the Department of Statistics at Rutgers University. Her research interests include spatial statistics, data assimilation, robust estimation.

04/06/2016

Bayesian Semiparametric Latent Variable Model: An Application on Fibroid Tumor Study

Mingan (Mike) Yang, PhD

Abstract: In parametric hierarchical models, it is standard practice to place mean and variance constraints on the latent variable distributions for the sake of identifiability and interpretability. Because incorporation of such constraints is challenging in semiparametric models that allow latent variable distributions to be unknown, previous methods either constrain the median or avoid constraints. In this article, we propose a centered stick-breaking process (CSBP), which induces mean and variance constraints on an unknown distribution in a hierarchical model. This is accomplished by viewing an unconstrained stick-breaking process as a parameter-expanded version of a CSBP. An efficient blocked Gibbs sampler is developed for approximate posterior computation. The methods are illustrated through a simulated example and an epidemiologic application.

Bio: Dr. Mingan Yang is an Assistant Professor of Biostatistics at graduate school of public health, San Diego State University. Upon graduation, he completed a postdoctoral research at Duke University and NIEHS, NIH, under the supervision of Dr. David Dunson. He specializes in Bayesian Statistics, Computational statistics, survival analysis, latent variable models, variable selection, and mixed effects models. He develops statistics methodology research with emphasis to address problems arising from health and medicine. Some research results are published in statistical journals such as Biometrics, Psychometrika, Computational Statistics & Data Analysis, and Biometrical Journal etc.

03/16/2016

Correlation and Mixture in High Dimensional Data: Should the Empirical Distribution Look Normal?

Armin Schwartzman, PhD

Abstract: Large scale multiple testing problems, such as in brain imaging and genomics, base their inference on a large number of z-scores. If most effects are null, it seems natural that the empirical distribution of z-scores should follow a standard normal distribution. But should it? In this talk Dr. Schwartzman will show two ways in which the empirical distribution of z-scores can be deceiving, because of correlation and mixture. First, following Efron’s (2007) conjecture, Dr. Schwartzman shows that even if the z-scores are standard normal, the empirical distribution may depart from it, due to strong correlation caused by hidden random effects. Instead, it may be approximated by a Gaussian mixture that generalizes Efron’s empirical null distribution. Second, Dr. Schwartzman shows that if the original data is a Gaussian mixture, then within-class standardization using a template-based EM algorithm produces z-scores whose empirical distribution looks standard normal. However, their true distribution has in fact lighter tails.

03/03/2016

Fence Methods for Genetic Application

Thuan Nguyen, PhD

Abstract: Model search strategies play an important role in finding simultaneous susceptibility genes that are associated with a trait. More particularly, model selection via the information criteria, such as the BIC with modifications, have received considerable attention in quantitative trait loci (QTL) mapping. However, such modifications often depend upon several factors, such as sample size, prior distribution, and the type of experiment, e.g., backcross, or intercross. These changes make it difficult to generalize the methods to all cases. The fence method avoids such limitations with a unified approach, and hence can be used more broadly. In this talk, the method is studied in the case of backcross experiments (BE). In particular, a variation of the fence, called restricted fence (RF), is applied to BE, and its performance is evaluated and compared with the existing methods. Furthermore, we incorporate our recently developed strategy for model selection with incomplete data, known as the E-MS algorithm, with the RF to address the common missing value concerns in BE. Our study reveals some interesting findings in association with the missing data mechanisms. The proposed method is illustrated with a real data analysis involving QTL mapping for an agricultural study on barley grains.

02/25/2016

Functional Response Models: A Unified Paradigm for Between- and Within-subject Attributes

Xin Tu, PhD

Abstract: Modern statistical methods provide a powerful tool to address complex statistical issues arising in clinical and translational research. However, the predominant statistical paradigm is only applicable to modeling relationships defined by within-subject attributes such as alcohol use and suicide from the same subject. Many relationships of interest in the age of the internet and mobile technology involve variables measuring between-subject attributes such as human interaction and such attributes are not amenable to treatment by conventional statistical models. In this talk, I will discuss a class of functional response models (FRM) to address this fundamental limitation in the current statistical paradigm. The between-subject attribute is not a concept unique to timely issues such as modeling human interaction in social networks, but is actually a fundamental barrier to understanding many classic statistical methods in order to extend them to address their limitations when applied to cutting-edge statistical problems in clinical and translational research. I will illustrate the FRM using a wide range of topics with both real and simulated data.

02/04/2016

Statistical Challenges of Using Electronic Data and Existing Research Infrastructure for CER

Mi-Ok Kim, PhD

Abstract: Developing the health information technology infrastructure to support comparative effectiveness research (CER) was a core objective of the American Recovery and Reinvestment Act of 2009. Many research networks, each including between 11,000 and 7.5 million patients each and more than 18 million in total, have established and numerous CER studies have been conducted. As compared to randomized clinical trials, these studies are less resource demanding and quickly collect data that are more representative of routine clinical care in large cohorts of patients over a long period of follow-up. Their utility, however, is restricted by the fact that treatment choice is affected by known or unknown prognostic factors, and consequently treatment groups are not directly comparable. This situation known as confounding by indication for treatment may render observational studies invalid and irrelevant unless properly addressed. Proper treatment of confounding is further complicated in data obtained from registries, network databases or the Electronic Health Record (EHR) where subjects or patients are commonly clustered in ways that may be relevant to the analysis. We will extend propensity score (PS) methodology and related sensitivity analysis to address measured and unmeasured confounding in the clustered data with the following aims:

Aim 1: Investigate how to optimally extend the PS methodology and identify what works best when

Aim 2: Develop a novel sensitivity analysis approach

Aim 3: Identify valid and most efficient PS methods for two existing CER studies.

We will use Monte Carlo computer simulation studies and real data including two existing CER studies. The real data examples will provide clinically plausible and interesting hierarchical data contexts and inform the design of the computer simulation studies about various types of outcomes that comprehend typical features of patient reported outcomes (PROs).