262 Section on Bayesian Statistical Science A.M. Roundtable Discussion (fee event)

Section on Bayesian Statistical Science
Tuesday, August 2, 7:00 a.m.–8:15 a.m.

Gender Issues in Academia and How to Balance Work and Life

✦Francesca Dominici, Harvard School of Public Health, Boston, 02115, *FDOMINIC@hsph.harvard.edu*

Key Words: gender issue, balancing work and life

Despite interventions by leaders in higher education, women are still under-represented in academic leadership positions. This dearth of women leaders is no longer a pipeline issue, raising questions as to the root causes for the persistence of this pattern. We have identified four themes as the root causes for the under-representation of women in leadership positions from focus group interviews of senior women faculty leaders at Johns Hopkins. These causes are found in routine practices surrounding leadership selection as well as in cultural assumptions about leadership potential and effectiveness. As part of this roundtable, I will discuss these findings and I will facilitate an informal discussion on how to balance work and life.

263 Section on Health Policy Statistics A.M. Roundtable Discussion (fee event)

Section on Health Policy Statistics
Tuesday, August 2, 7:00 a.m.–8:15 a.m.

Do You Want To Visit Australasia?

✦Louise Ryan, CSIRO, Australia, North Ryde, Sydney, _ 1670 Australia, *Louise.Ryan@csiro.au*

Key Words: Australia, New Zealand, Australasia

Have you ever thought about visiting the Australasian region for a short research visit or sabattical, to attend a conference, or to take up short-term or long-term employment? During this session we will explore what it's like living and working in Australasia, the differences in statistical projects between the American and Australasian regions, and what job opportunities are currently available. This session will be led by Louise Ryan, Chief of Mathematics, Informatics and Statistics at CSIRO. The Commonwealth Scientific and Industrial Research Organisation is Australia's national science agency and one of the largest and most diverse research agencies in the world (housing approximately 6500 employees).

264 Section on Physical and Engineering Sciences A.M. Roundtable Discussion (fee event)

Section on Physical and Engineering Sciences
Tuesday, August 2, 7:00 a.m.–8:15 a.m.

Engaging Stochastic Spatiotemporal Methodologies in Renewable Energy Research

✦Alexander Kolovos, SAS Institute, Inc., 100 SAS Campus Dr., S3042, Cary, NC 27513 USA, *Alexander.Kolovos@sas.com*

Key Words: spatiotemporal, stochastic modeling, renewable energy, energy research, solar, wind

This roundtable builds on a discussion initialized at JSM2010. Last year the panel explored the interest in connecting statistical methodologies with research in the fields of renewable energy and sustainability. In particular, stochastic spatiotemporal analysis can provide a plethora of tools for fundamental aspects in the modeling of attributes related to renewable energy resources such as solar radiation, wind fields, tidal waves. In alignment with this year's JSM all-encompassing theme, the goal now is to take a further step forward and engage energy-related disciplines in the industry and the academia to benefit their research from the availability of advanced methodologies in space-time analysis. Ideally, the session looks into enabling a communication core between statisticians and specialists in renewable energy and sustainability to bring forward the potential and benefits of shared research in these disciplines.

265 Section on Quality and Productivity A.M. Roundtable Discussion (fee event)

Section on Quality and Productivity
Tuesday, August 2, 7:00 a.m.–8:15 a.m.

Effective Statistical Training in Industry

✦Willis Jensen, W. L. Gore & Associates, Inc., 3750 West Kiltie Lane, Flagstaff, AZ 86003, *wjensen@wlgore.com*

Key Words: Training Objectives, Training Myths, Teaching Techniques, Socratic Method

Many statisticians in industry have responsibilities to develop and deliver statistical training to different audiences, including engineers, scientists and business leaders. We'll talk about how we can make this training more effective including some common training myths to recognize. Some key questions that we'll discuss are: 1 - What are appropriate objectives for training? 2 - Based on the objectives, what training topics should be covered in an industrial training? 3 - What are good training techniques/methods/ approaches? 4 - How do you evaluate the effectiveness of training?

266 Section on Statistical Computing A.M. Roundtable Discussion (fee event)

Section on Statistical Computing
Tuesday, August 2, 7:00 a.m.–8:15 a.m.

Introductory Data Analysis Courses With (And Without) R

✦John Emerson, Yale University, P.O. Box 208290, New Haven, CT 06520-8290 USA, *john.emerson@yale.edu*

Key Words: introductory statistics, data analysis, computing, software, education

This roundtable focuses on introductory data analysis courses intended primarily for undergraduate students. Such a course can be an alternative to a traditional introductory statistics course, but is more likely to serve as a second course in statistics. The roundtable will provide a forum for discussion of existing or planned courses in this area, including the choice of computing platforms.

267 Section on Statistical Consulting A.M. Roundtable Discussion (fee event)

Section on Statistical Consulting
Tuesday, August 2, 7:00 a.m.–8:15 a.m.

Statistical Consulting As A Career Path

✦Carlos Alzola, Data Insights Inc., , *calzola@verizon.net*;
✦Anamaria Segnini Kazanis, ASKSTATS Consulting LLC, 24284 Woodham Rd., Novi, MI 48374 USA, *akazanis@umich.edu*

Key Words: Statistical Consulting, Statistical Career, Statistical Career Path, Statistical Apprenticeship, Statistical Internships

Statistics is a field rife with opportunity. Statistics professionals and expertise are needed in all sorts of fields and industries. Statisticians are hired by government agencies (Census Bureau, Federal Reserve, FDA, NIH, NIST to name a few), pharmaceuticals, financial, market research companies among other. They work in fields as diverse as economics, survey research, quality control in industrial processes and even study human rights violations. A statistical consultant is expected to be an expert in his field and also to be intimately familiar with the industry in which he/she is consulting. For that reason, entry level positions are hard to come by. Initial apprenticeship usually starts during graduate school in university sponsored centers for statistical consultation and/or college internships. An aspiring consulting statistician needs to develop an area of expertise in addition to his/her statistical training that will be of interest to a potential client. This round table will discuss opportunities and options for the young (and and not so young)statistics professional in consulting.

268 Section on Statistical Education (fee event)

Section on Statistical Education
Tuesday, August 2, 7:00 a.m.–8:15 a.m.

Teach Statistical Literacy With Epidemiology!

✦Daniel T Kaplan, Macalester College, 1600 Grand Ave., Saint Paul, MN 55105 USA, *kaplan@macalester.edu*

Key Words: statistical literacy, epidemiology, education, undergraduate

For the last two years, I have been teaching an undergraduate epidemiology course at Macalester College. It has no pre-requisites and, from the very beginning, was popular with students. Epidemiology provides a wonderful setting for teaching many important concepts relating to statistical literacy. The setting is very compelling to students and helps to inform both their professional choices and future personal healthcare choices. Most of the issues that arise in statistical literacy courses are important in epidemiology, but it has a much more grounded notion of causation. In this round-table, I'll present the outline of the course, some of the quantitative and statistical activities we do, and the "Epidemiology 101" recommendations in the CCAS "Consensus Report on Public Health and Undergraduate Education." Then we can discuss how epidemiology might fit in at your own institution.

Using R In Introductory Statistics Courses

✦Amy Wagaman, Amherst College, 01002, *awagaman@amherst.edu*

Key Words: R, introductory statistics

This roundtable will discuss the benefits and challenges of using R as the statistical computing software in introductory statistics courses. Examples of user-friendly GUIs including R Commander and Rattle will be discussed as well as activities for students and student reactions to the software in classroom environments with and without the GUIs.

269 Section on Statistics and the Environment A.M. Roundtable Discussion (fee event)

Section on Statistics and the Environment
Tuesday, August 2, 7:00 a.m.–8:15 a.m.

Statistics For Wind Energy

✦Marc G. Genton, Texas A&M University, Department of Statistics, TAMU, College Station, TX 77843-3143 USA, *genton@stat.tamu.edu*

Key Words: Energy, Forecast, Spatial statistics, Time series, Wind

Global large scale penetration of wind energy is accompanied by significant challenges due to the intermittent and unstable nature of wind. High quality short-term wind speed forecasting is critical to reliable

and secure power system operations. We will discuss how statistical techniques can enable wind energy to be more efficiently incorporated into the electrical grid.

270 Section on Teaching of Statistics in the Health Sciences (fee event)

Section on Teaching of Statistics in the Health Sciences

Tuesday, August 2, 7:00 a.m.–8:15 a.m.

A Program To Enhance The Collaboration, Education And Communication Skills Of Biostatistics Graduate Students

✦Miranda E. Kroehl, University of Colorado Denver, Aurora, CO 80045, *miranda.kroehl@ucdenver.edu*; Michael Jacobson, University of Colorado Denver; Katie Den Ouden, University of Colorado Denver; Carole Basile, University of Colorado Denver

Key Words: graduate program, communication skills

Graduate students in biostatistics must be prepared with the necessary skills to face career challenges of the 21st century. In addition to research and analysis competencies, students must be able to communicate their work not only to other statisticians, but to a variety of audiences. The National Science Foundation (NSF) recognizes that communication is not normally emphasized in traditional science/technology/engineering/math (STEM) graduate programs. NSF developed the Graduate STEM Fellows in K-12 Education (GK-12) program to enhance the communication skills of graduate students through fellowship programs. The University of Colorado Denver GK-12 Transforming Experiences Project works to enhance communication skills of its fellowship recipients through a series of workshops and communication evaluations. This roundtable will discuss how these types of activities can be incorporated into biostatistics graduate programs to help train statistics students to become better communicators, educators and collaborators. We will encourage a group discussion of these activities and are interested to learn about any communication enhancement activities implemented in other programs.

271 Introductory Overview Lecture: Statistical Graphics

ASA

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Statistical Graphics

✦Dianne Cook, Iowa State University, 2415 Snedecor Hall, Ames, IA 50011 USA, *dicook@iastate.edu*

This talk will provide an introduction to graphics, as used in statistical analyses. Examples will be taken from topics such as the US elections, US air traffic, real estate crisis, clustering of music clips, classification of chocolates, and the the environmental effects of the BP oil spill. All of the graphs shown will be reproducible in R, and some mix of static and interactive graphics will be used. Emphasis will be placed on the role graphics plays in statistical analysis, why, where, when and how to display information in graphical form.

272 ASA College Stat Bowl I

ASA, ENAR, WNAR, IMS, International Chinese Statistical Association, International Indian Statistical Association, SSC

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

273 New techniques for functional data analysis

Section on Nonparametric Statistics, International Chinese Statistical Association, International Indian Statistical Association, SSC, Section on Statistical Computing

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Achieving Near-Perfect Classification for Functional Data

✦Peter Hall, The University of Melbourne and UC Davis, , *halpstat@ms.unimelb.edu.au*; Aurore Delaigle, University of Melbourne

It can be shown that, in functional data classification problems, perfect asymptotic classification is possible, making use of the intrinsic very high dimensional nature of functional data. This performance is often achieved by linear methods, which are optimal in important cases. These results point to a marked difference between classification for functional data and its counterpart in conventional multivariate analysis, where dimension is kept fixed as sample size diverges. In the latter setting, linear methods can sometimes be quite inefficient, and there are no prospects for asymptotically perfect classification, except in pathological cases where, for example, a variance vanishes. By way of contrast, in finite samples of functional data, good performance can be achieved by truncated versions of linear methods. Truncation can be implemented by partial least-squares or projection onto a finite number of principal components, using, in both cases, cross-validation to determine the truncation point.

Nonlinear Functional Regression

✦Gareth James, University of Southern California, IOM Department, 401 Bridge Hall, Los Angeles, CA 90089-0809, *gareth@usc.edu*

Key Words: Functional Regression, Nonlinear, Penalized Regression

We suggest a new method, called "Functional Additive Regression", or FAR, for efficiently performing high dimensional functional regression. FAR extends the usual linear regression model involving a functional predictor, X(t), and a scalar response, Y, in two key respects. First, FAR uses a penalized least squares optimization approach to efficiently deal with high dimensional problems involving a large number of different functional predictors. Second, FAR extends beyond the standard linear regression setting to fit general non-linear additive models. We demonstrate that FAR can be implemented with a wide

range of penalty functions using a highly efficient coordinate descent algorithm. Theoretical results are developed which provide motivation for the FAR optimization criterion. Finally, we show through simulations and a real data set that FAR can significantly outperform competing methods. This is joint work with Yingying Fan.

Additive Modeling of Functional Gradients

✦Fang Yao, University of Toronto, Department of Statistics, 100 St. George Street, Toronto, ON M5S3G3 Canada, *fyao@utstat.toronto.edu*; Hans-Georg Mueller, University of California, Davis

Key Words: Functinal Derivative, Functional Data, Functional Regression, Gradient Field, Nonparametric Differentiation, Principal Components

We consider the problem of estimating functional derivatives and gradients in the framework of a functional regression setting where one observes functional predictors and scalar responses. Derivatives are then defined as functional directional derivatives which indicate how changes in the predictor function in a specified functional direction are associated with corresponding changes in the scalar response. Aiming at a model-free approach, navigating the curse of dimensionality requires the imposition of suitable structural constraints. Accordingly, we develop functional derivative estimation within an additive regression framework. This approach requires nothing more than estimating derivatives of one-dimensional nonparametric regressions, and thus is computationally very straightforward to implement, while it also provides substantial flexibility, fast computation and is consistent. We demonstrate the consistent estimation and interpretation of the resulting functional derivatives and functional gradient fields in a study of the dependence of lifetime fertility of flies on early life reproductive trajectories.

274 Some Recent Developments in inference for mixture and linear errors-in-variables models

International Indian Statistical Association, International Chinese Statistical Association, International Indian Statistical Association, SSC

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

L2E Estimation for Finite Mixture of Regression Models with Applications

✦T. N. Sriram, University of Georgia, Department of Statistics, Statistics Building, Cedar Street, Athens, GA 30602, *tn@stat.uga.edu*

Key Words: Asymptotic normality, Consistency, Extreme values, L2E functional, Robustness

For count response, we propose a robust estimation method for finite mixture of regression models based on minimum integrated L2 distance between parametric conditional and true conditional mass functions. The estimator, called the L2E, is shown to be consistent and asymptotically normal. We demonstrate the efficiency and robustness of the L2E estimator through a variety of simulation studies. Finally, the L2E in conjunction with a two-component Poisson mixture regression model is shown to provide a competitive fit to a hospital length of stay data, which contains extreme values. Our L2E approach is also shown to readily extend to the continuous response case, thereby expanding its scope of use and application.

Testing the Number of Components in a Semiparametric Mixture Model via Empirical Likelihood

✦Lianfen Qian, Florida Atlantic University, 777 Glades Road, Department of Mathematical Sciences, Boca Raton, FL 33431, *Lqian@fau.edu*

Key Words: Number of Components, Mixture Model, Empirical Likelihood

This talk first proposes a partial empirical likelihood ratio (PELR) statistic to test for 1-component against k-component (k>1) in a semiparametric mixture model. We will examine the distribution of the PELR statistic under the null. Based on the PELR statistic we will propose an empirical likelihood based estimator of the number of components. Simulation studies will be reported on the distributional properties of the proposed test statistic under null hypothesis and the performance of the estimation for finite sample size.

Mixture Models and High-Dimensional Data

✦Soumendra Nath Lahiri, ASA, Department of Statistics; MS 3143, Texas A & M Univ, College Station, 77843, *snlahiri@stat.tamu.edu*; Subhodeep Mukhopadhyay, Texas A & M University

Key Words: High dimensional data, Mixture models, optimal classifier

In this talk, we consider Gaussian mixture models in high dimensional set up where the dimension of the observations diverges with the sample size. We derive asymptotic properties of the optimal classifier based on mixture models and illustrate the results with applications to Gene expression data.

Goodness-of-Fit Test in Linear Errors-in-Variables Models

✦Weixing Song, Kansas State University, , *weixing@ksu.edu*

Key Words: Lack-if-Fit Test, Bootstrap Approximation, L2 Distance

A class of Bickel-Rosenblatt type goodness-of-fit tests is proposed for fitting a parametric family to the regression error density function in linear errors-in-variables models. These tests are based on a class of L2 distances between a kernel density estimator of the residual and an estimator of its expectation under null hypothesis. The paper investigates asymptotic normality of the null distribution of the proposed test statistics. Asymptotic power of these tests under certain fixed and local alternatives is also considered, and an optimal test within the class is identified. A parametric bootstrap algorithm is proposed to implement the proposed test procedure when the sample size is small or moderate. A finite sample simulation study shows very desirable finite sample behavior of the proposed inference procedures.

275 Statistical modeling of genetic sequence data ●

ENAR, International Chinese Statistical Association, International Indian Statistical Association, Section on Statistics in Epidemiology

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Genetic Association Studies Using Individual and Pooled Sequence Data

✦Hongyu Zhao, Yale University, 300 George Street, New Haven, CT , *hongyu.zhao@yale.edu*; Joon Sang Lee, Yale University

Key Words: sequencing, rare variants, pooling, association, complex traits, statistical genomics

Second generation sequencing technologies have great promises to identify and associated novel DNA variants with complex diseases. Because the data are extremely information rich yet still expensive to collect, the design and analysis of such data presents many statistical challenges. In this talk, we will discuss several issues encountered in our ongoing studies and describe how statistical principles can be applied to improve design efficiency and derive more robust and efficient statistical tests.

Statistical Issues in the Analysis of Disease Association with Next-Generation Sequence Data

✦Danyu Lin, University of North Carolina, , *lin@bios.unc.edu*

Key Words: next-generation sequencing, rare variants, copy number variants, single nucleotide polymorphisms, quantitative traits, selective genotyping

Next-generation sequencing technologies are being employed in an increasing number of genetic associations studies. There are many unique statistical challenges in the designs and analysis of such studies. In this talk, I will present an overview of these challenges, focusing on three different issues: (1) analysis of disease association with rare variants; (2) analysis of disease association with copy number variations; (3) analysis of quantitative traits with outcome-dependent sampling. There are strong motivations for addressing these three issues: (1) because we are particularly interested in rare variants detected by next-generation sequencing technologies, standard statistical methods based on asymptotic distributions are inappropriate; (2) the copy number variations are not directly measured, and the intensity data from next-generation sequencing technologies are quite different from genome-wide SNP array data; (3) genotyping individuals with extreme phenotype values can be very cost-effective for large cohort studies. We will describe our solutions to these problems and illustrate them with simulated and real data.

Replication and Analysis of Complex Trait Rare Variant Association Studies

✦Suzanne Margaret Leal, Baylor College of Medicine, One Baylor Plaza, 700D, Houston, TX 77030 USA, *sleal@bcm.edu*

Key Words: rare variants, sequence data, association analysis, replication, genetics

There is overwhelming evidence that rare variants play an important role in complex disease etiology. Currently many studies are generating exome sequence data using next generation sequencing in order to detect rare variant complex trait associations. Association methods used to analyze common variants should not be used for the analysis or rare variants since they are grossly underpowered. Instead methods which have been developed to test for association with rare variants which rely on aggregating information on rare variants within a region, e.g. gene should be used. Once associations are detected it is important to replicate the findings in an independent sample, since findings can be false positives even if the family wise error rate is well controlled. Spurious associations can occur due to improperly controlling for population admixture/substructure and sequencing errors. Methods to analyze rare variant association data for complex traits will be presented and their power will be compared for a variety of underlying genetic models. Additionally strategies for gene-based replication will be compared

Statistical Analysis Approaches to Sequence-Based Association Studies

✦Nicholas Schork, The Scripps Research Institute, 3344 North Torrey Pines Court, Suite 300, La Jolla, CA 92037 USA, *nschork@scripps.edu*

Key Words: DNA sequence, distance analysis, regression models, polymorphism, mutation

The limitations of genome-wide association (GWA) studies that focus on the phenotypic influence of common genetic variants have motivated human geneticists to consider the contribution of rare variants to phenotypic expression. The increasing availability of high-throughput sequencing technology has enabled studies of rare variants, but will not be sufficient for their success since appropriate analytical methods are also needed. We evaluate the intuitions and modeling constructs behind many statistical analysis approaches to testing associations between a phenotype and collections of rare variants in a defined genomic region or set of regions. We also apply these methods to actual sequence data in an effort to showcase their utility and limitations. Ultimately, although a wide variety of analytical approaches exist, more work is needed to refine them and determine their properties and power in different contexts.

276 Partial Identification and Inference in Nonlinear Models ■

JBES-Journal of Business & Economic Statistics, International Chinese Statistical Association, International Indian Statistical Association

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Identification of Regressions with Missing Covariate Data

✦Federico Bugni, Duke University, 213 Social Science Building, Durham, NC 27708 USA, *federico.bugni@duke.edu*; Joseph Hotz, Duke University ; Esteban Aucejo, Duke University

Key Words: missing data, outer identified sets, sharp sets

This paper examines the problem of identification and inference on parametric models when there are missing data, with special focus on the case when covariates, denoted by X, are missing. Our econometric model is given by a conditional moment condition implied by the assumption that X is strictly exogenous. At the same time, we assume that the distribution of the missing data is unknown. We confront the missing data problem by adopting a worst case scenario approach. We characterize the sharp identified set and argue that this set is usually prohibitively complex to compute or to use for inference. Given this difficulty, we consider the construction of outer identified sets (that is, supersets of the identified set) that are easier to compute and can still provide a characterization of the parameter of interest. Two different outer identification strategies are discussed. Both of these strategies are shown to contain non-trivial identifying power and are relatively easy to compute and to be used for inference.

Asymptotic Distortions in Locally Misspecified Moment Inequality Models

Federico Bugni, Duke University ; ✦Ivan Canay, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208 USA, *iacany@northwestern.edu*; Patrik Guggenberger, University of California at San Diego

Key Words: local misspecification, size distortion, confidence sets

This paper studies the behavior under local misspecification of several confidence sets (CSs) commonly used in the literature on inference in moment inequality models. We suggest the degree of asymptotic confidence size distortion as an alternative criterion to power to choose among competing inference methods, and apply this criterion to compare across critical values and test statistics employed in the construction of CSs. We find two important results under weak assumptions. First, we show that CSs based on subsampling and generalized moment selection (GMS, Andrews and Soares (2010)) suffer from the same degree of asymptotic confidence size distortion, despite the fact that the latter can lead to strictly smaller CSs under correct model specification. Second, we show that CSs based on the quasi-likelihood ratio test statistic have asymptotic confidence size that can be an arbitrary fraction of the asymptotic confidence size of CSs obtained by using the modified method of moments. Our results are supported by Monte Carlo simulations.

Sharp Identifcation Regions in Models with Convex Moment Predictions

✦Francesca Molinari, Cornell University, 492 Uris Hall , Ithica, NY 14853 USA, *fm72@cornell.edu*; Arie Beresteanu, University of Pittsburgh; Ilya Molchanov, University of Bern

Key Words: Partial Identification, Random Sets, Aumann expectation

We provide a tractable characterization of the sharp identification region of the parameters in a broad class of incomplete econometric models. Models in this class have set valued predictions that yield a convex set of conditional or unconditional moments for the observable model variables. In short, we call these models with convex moment predictions. Examples include static, simultaneous move finite games of complete and incomplete information in the presence of multiple equilibria; best linear predictors with interval outcome and covariate data; and random utility models of multinomial choice in the presence of interval regressors data. Given a candidate value for ; we establish that the convex set of moments yielded by the model predictions can be represented as the Aumann expectation of a properly defined random set. The sharp identification region can then be obtained as the set of minimizers of the distance from a properly specified vector of moments of random variables to this Aumann expectation. Algorithms in convex programming can be exploited to efficiently verify whether a candidate is in the sharp identification region.

Inference on Panel Data Models with Endogenous Censoring

✦Elie Tamer, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208 USA, *tamer@northwestern.edu*; Shakeeb Khan, Duke University; Maria Ponomareva, University of Western Ontario

Key Words: moment inequality, panel data, sharp bounds

This paper explores the identifiability of regression coefficients in a linear panel data models with endogenous censoring. We take an approach that is based on deriving conditional moment inequalities which can be used to infer the identified regions of the parameter space. We show that this set of moment inequalities is complete: the bounds we attain for the parameters based on these inequalities are sharp- they constitute the smallest subset of the parameter space that is consistent with the assumptions of the model and the data. We consider two separate set of assumptions, each controlling for unobserved heterogeneity with an individual specific fixed effect. The first imposes a strict stationarity assumption on the unobserved disturbance terms, along the lines of Manski(1987), Honore(1993). Under such a condition we propose a maximum score based procedure which consistently estimates the sharp set. The second set of assumptions relaxes the stationarity condition but imposes cross-sectional homoskedasticity. Based on these alternative conditions we propose a differenced maximum rank procedure. An inference procedure for both models based on subsampling is proposed.

277 Recent developments in addressing multiplicity issues in clinical trials ■●

Biopharmaceutical Section, Section on Health Policy Statistics, Scientific and Public Affairs Advisory Committee, Section for Statistical Programmers and Analysts

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Graphical Approaches to Multiple Test Procedures

✦Frank Bretz, Novartis Pharma AG, Lichtstr. 35, Basel, International 4002 Switzerland, *frank.bretz@novartis.com*

Key Words: gatekeeping, fallback, Bonferroni, truncated, Simes, parametric

A variety of sequentially rejective, weighted Bonferroni tests have been proposed for clinical trials with multiple treatment arms or endpoints, such as gatekeeping procedures, fixed sequence tests and fallback procedures. These approaches allow one to reflect the difference in importance as well as the relationship between the various research questions

by choosing suitable weights. In this presentation we propose a simple iterative graphical approach to construct and perform such Bonferroni-type tests. The resulting multiple test procedures are expressed as directed, weighted graphs, where each node corresponds to an elementary null hypothesis, together with a simple algorithm to generate such graphs while sequentially testing the individual hypotheses. A case study is used to demonstrate how this approach can be used to tailor a multiple test procedure to given study objectives. The approach is also illustrated with the visualization of several common gatekeeping strategies. Extensions of the basic graphical approach are discussed briefly, including parametric methods accounting for the correlation between the test statistics, weighted Simes tests and truncated test procedures.

Weighted Multiple Comparisons Procedures for Clinical Trials

✦Brian L Wiens, Alcon Laboratories, Inc., 6201 South Freeway, TC 41, Fort Worth, TX 76013 USA, *brian.wiens@alconlabs.com*; Alex Dmitrienko, Eli Lilly and Company

Key Words: gatekeeping, familywise error rate, optimality

In choosing a multiple comparison procedure (MCP) for a clinical trial, the weights assigned to each hypothesis must be considered. Optimal weighting depends on the definition of optimality: maximizing the probability of rejecting at least one hypothesis or of rejecting all hypotheses, equalizing the probability of rejecting each hypothesis or some other criterion. We consider various optimality criteria in comparing several MCPs, including the Bonferroni, Holm, fallback and 4A procedures. We find weights that maximize the difference in various pairs of tests in special situations. Choosing weights to equalize the power of various hypotheses makes sense primarily when using the Bonferroni procedure, which is useful in parallel gatekeeping but suboptimal for a single family. We close with some recommendations.

Use of Prior Information for Improving the Power of Multiple Testing Procedures

✦Mohammad Huque, FDA, 10903 New Hampshire Ave, Bldg. 21, room, Silver Spring, MD 20993, *Mohammad.Huque@fda.hhs.gov*; Mohamed A. Alosh, FDA

Key Words: Multiplicity, adaptive allocation, prior information

A clinical trial might involve more than one clinically relevant endpoint each of which is sufficient on its own to characterize the treatment benefit, yet due to power consideration one of these endpoints might be given a lower priority in a sequential testing for establishing an efficacy claim. The same situation arises when testing for a pre-specified clinically 'important' component of a composite endpoint following testing for the composite endpoint, or testing for pre-specified subgroup after testing for the total population. Various multiple testing procedures have been proposed in the literature with different allocation of the type I error rates between the two endpoints and possibly allowing the significant level for testing the second endpoint to adapt to the findings of the first endpoint. However, it is well-recognized in the literature that no single method outperforms other methods. In this presentation we investigate the utility of using additional information, which might be available at the design stage, in comparing the power performance of different multiplicity procedures for rejecting at least one null hypothesis as the criteria for a positive trial.

Testing Individual Hypothesis Marginally at 0.05 Without Inflation of the Family-Wise Error Rate

✦David Li, Pfizer, , *david.li1@pfizer.com*

Key Words: multiple testing, consonant tests, collective evidence, consistency

As a clinical trial statistician, I was challenged numerous times by the question "can we test each hypothesis at 0.05?" Being asked again and again, finally I started asking myself: Can we test each hypothesis at 0.05 and how? This presentation will provide an answer to the question by constructing consonant multiple testing procedures. The consonant requirement is relatively inconsequential in a clinical trial when multiple primary endpoints have low correlations (rho = 0.3). For other scenarios, a modified version is proposed. These procedures are particularly useful when the collective evidence across endpoints or consistency across endpoints is expected or required.

278 Improving physician understanding of biostatistics in the medical literature

■●

Section on Teaching of Statistics in the Health Sciences, Section on Statistical Education, Section on Health Policy Statistics, Section on Statistics in Epidemiology

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Medicine Residents' Understanding of Biostatistics and Results in the Medical Literature

✦Donna Windish, Yale University School of Medicine, Waterbury Hospital, Waterbury, CT 06708, *donna.windish@yale.edu*

Key Words: medical statistics, survey, reading comprehension

Objective To evaluate residents' understanding of biostatistics and interpretation of research results. Methods: Multi-program cross-sectional survey of internal medicine residents. Outcome: Percentage of questions correct on a biostatistics multiple-choice knowledge test. Results: 277 of 367 residents (76%) in 11 programs participated. The mean percentage correct on statistical knowledge and interpretation of results was 41% vs 72% for fellows and general medicine faculty with research training (P< .001). Higher scores in residents were associated with additional advanced degrees (50% vs 40%; P< .001); prior biostatistics training (45% vs 38%; P=.001); enrollment in a university-based training program (43% vs 36%; P=.002); and male sex (44% vs 39%; P=.004). On individual knowledge questions, 82% correctly interpreted a relative risk. Residents were less likely to know how to interpret an adjusted odds ratio from a multivariate regression analysis (37%) or the results of a Kaplan-Meier analysis (11%). While 75% indicated they did not understand all of the statistics in journal articles, 95% felt it was important to understand these concepts to be an intelligent reader

Promoting Clinical Statistics Literacy of Emergency Medicine Residents with Clicker Technology

✦Penny S Reynolds, Virginia Commonwealth University Medical Center, Department of Emergency Medicine, MCV Campus, Richmond, VA 23298 USA, *psreynolds@vcu.edu*

Key Words: Emergency Medicine, in-class assessment tools, clinical statistics, comprehension

Objective: The goal was to increase resident reading comprehension of clinical statistics over 10 months. Methods: Emergency Medicine residents were given lectures covering 13 learning objectives in 3 concept areas in clinical statistics identified as competency requirements. Lecture content de-emphasized computations in favor of concepts. In-class assessment tools ("clickers") provided immediate feedback. Supplemental readings were posted online before the lecture. Assessment was based on pre- and post-test survey scores. Results: Reading comprehension of basic concepts remained poor. Median pre-test score was 30%; post-test was 35%. There was no change in knowledge of design, power, sensitivity and specificity, and bias. Understanding of risk increased from 25% to 70%; paradoxically, residents were unable to compute related metrics e.g. number needed to treat. On-line tracking showed that few residents availed themselves of supplementary resources.Conclusion: Physicians must be able to stay current with the vast and rapidly expanding body of medical research literature. However, resident motivation is lacking. Clickers increased engagement, but did not reinforce learning

Medical Students and Statistics: Challenges in Teaching, Learning, and Assessment

✦Philip M Sedgwick, Centre for Medical and Healthcare Education,, St. George's, University of London, London, SW17 0RE UK, *p.sedgwick@sgul.ac.uk*

Key Words: Medical statistics, Integrated teaching, Contextualised learning

Medical statistics is a core component of the medical curriculum in the UK. In today's era of evidence based medicine, doctors need to assimilate knowledge and evidence for their practice based on literature that incorporates an increasing use of statistics. However, many students still fail to see the relevance of statistics to clinical practice. Summary of work: This presentation will describe the teaching and learning of statistics in a UK medical school. The content and delivery of the curriculum reflects the importance of teaching statistical thinking rather than statistical techniques. By integrating teaching with other disciplines, in particular clinical communication, students' learning can be contextualised. Emphasis is placed on a learning process rather than gaining knowledge, incorporating the ability to evaluate data as well as developing skills to interact with patients and colleagues. A variety of teaching methods will be described including videos, small group exercises plus self-assessment statistical exercises. Assessment has been developed to reflect approaches to teaching and learning.

279 Analytical Challenges and Best Statistical Practices in Smoking Cessation Research ■●

Section on Statistics in Epidemiology, ENAR, International Indian Statistical Association, Section on Health Policy Statistics

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Analysis of Longitudinal Smoking Outcomes with Missing Data Under a Simple Nonignorable Multiple Imputation Model

✦Donald Hedeker, University of Illinois at Chicago, School of Public Health (MC 923), 1603 W. Taylor St., room 955, Chicago, IL 60612-4336, *hedeker@uic.edu*; Hakan Demirtas, University of Illinois at Chicago; Robin J. Mermelstein, University of Illinois at Chicago

Key Words: missing data, binary outcomes, multiple imputation, longitudinal data

We consider the problem of missing smoking outcomes in a longitudinal two-group design. In this situation, some researchers deterministically recode the missing data to smoking, or assume that the missing data are equal to the last available smoking outcome for a subject (i.e., last observation carried forward, or LOCF). Building on our work for data at a single timepoint (Hedeker, Mermelstein, & Demirtas, 2007, Addiction), we describe how these deterministic assumptions can be relaxed by allowing missingness to be imperfectly related to the smoking outcome, and stratified on past patterns of the smoking outcome. Thus, one can examine the robustness of study findings to the assumed strength of the relationship between missingness and smoking. Our approach uses multiple imputation to take into account the uncertainty inherent in the imputed data. We illustrate the methods using data from a smoking cessation study, and describe computer syntax to perform the analyses.

Model-Based Analysis of Heaped Longitudinal Cigarette Count Data in Smoking Cessation Trials

✦Daniel Heitjan, University of Pennsylvania, 19104, *dheitjan@upenn.edu*; Sandra D. Griffith, University of Pennsylvania; Yimei Li, Children's Hospital of Philadelphia; Hao Wang, Johns Hopkins University; E. Paul Wileyto, University of Pennsylvania

Key Words: smoking cessation, longitudinal, software, imputation, count data, calibration

In smoking cessation trials it is common to collect daily cigarette counts, although primary analyses use coarser outcomes such as smoking status at end of treatment. Proper modeling of the daily counts could give more efficient and detailed analyses, allowing the estimation of treatment effects on mean cigarette consumption in non-quitters and the time dependence of treatment effects. Unfortunately, cigarette count data are often heaped, in the sense of being reported rounded to multiples of five, ten, or twenty. Heaping can substantially bias analyses of mean count, an intractable problem unless one has a valid model to predict heaped from true counts. We are in possession of a dataset where cigarette counts were measured by both electronic diaries (not subject to heaping) and conventional recall (strongly heaped). We will

use these data to create a model to impute accurate cigarette counts from the recall data in a clinical trial. We will fit zero-inflated longitudinal Poisson models to the imputed accurate data, implementing the procedure in a mixture of R and SAS software.

Analyzing Multivariate Longitudinal Data from Smoking Cessation Studies

✦Joel A Dubin, University of Waterloo, Waterloo, ON Canada, *jdubin@uwaterloo.ca*; Jesse D. Raffa, University of Waterloo

Key Words: abstinence success, adverse events, compliance, dose-ranging, pharmacotherapy, smoothing

Since obtaining several longitudinal measures from smoking cessation studies is quite common (e.g., daily and/or weekly cigarette consumption / smoking status, daily/weekly alcohol use, daily/weekly adverse event experience, treatment compliance information, other time-varying covariates such as weight change, etc.), there is interest in analyzing these data collectively, without much loss of information. I will discuss different strategies for incorporating this data into the analysis of smoking cessation studies, using a smoking cessation pharmacotherapy trial as the primary example dataset. A key analysis issue is that not all the longitudinal measurements are observed on the same time points.

A Longitudinal Social Network Analysis of Tobacco Use and Friendship Dynamics Among First-Year College Students

✦Stephanie R Land, University of Pittsburgh, 201 N. Craig Street, Suite 350, Pittsburgh, PA 15217, *land@pitt.edu*; Ju-Sung Lee, Carnegie Mellon University; Kristina Cooper, University of Pittsburgh

Key Words: tobacco, social networks, longitudinal data

Recent innovations in the modeling of social networks have yielded insights into the development of leading causes of mortality, such as tobacco use and obesity. This methodology has the potential to address important questions regarding the development of social networks, and the propagation of behaviors within social networks. Our study focuses on tobacco smoking with cigarettes and hookahs in the context of social networks among first-year college students. Participants each named up to 10 social contacts, many of whom were also participants, and provided tobacco use and social network data for themselves and their contacts in longitudinal assessments. The social network is densely connected, not limited to ego networks. We will present social network graphs and the longitudinal associations of tobacco behaviors with friendship dynamics, network popularity, betweenness, and eigenvector centrality. Ours is one of the largest longitudinal social network datasets available, providing rich opportunities for method development.

280 Statewide Education Longitudinal Data Systems: A Federal/State Partnership to Support Data-Driven Decision Making

■●

Section on Government Statistics, Social Statistics Section, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Statewide Longitudinal Data Systems: The Federal Role in Supporting SLDS Development

✦Andrew White, National Center for Education Statistics, US Dept of ED, 1990 K Street NW, NCES, 9th floor, Washington, DC 20006, *Andrew.White@ed.gov*; Jack Buckley, National Center for Education Statistics

Key Words: data-driven decision making, statewide longitudinal data systems, student education records

Federal funds authorized by the Educational Technical Assistance Act and the America Recovery and Reinvestment Act were provided through grants to states to assist in the development and implementation of Statewide Longitudinal Data Systems (SLDS) based on student education records. These systems are expected to increase States' efficient management, analysis, and use of education data to improve student learning. These systems are also expected to support research to increase student achievement and close achievement gaps. Although the initial SLDS support was for systems in grades K-12, the support has expanded to bridge across the pre-K, elementary/secondary, postsecondary, and workforce continuum. This effort is supported by a Federal State partnership in which the Federal role involves, providing funding, technical advise, and coordination in support of the individual state activities. This paper describes the federal role in this effort and discusses examples of the ways the resulting longitudinal data can be used.

Statewide Education Longitudinal Data Systems: Privacy and Data Protection Plans

✦Kathleen Styles, U.S. Department of Education, Room 2W332, 400 Maryland Avenue, SW, Washington, 20202, *Kathleen.Styles@ed.gov*; Marilyn Seastrom, National Center for Education Statistics

Key Words: Privacy, de-identification, anonymization, data protections, data sharing, data use agreements

The development of Statewide Longitudinal Data Systems (SLDS) leads to an expansion in the amount of personally identifiable information (PII) that is maintained on America's students. For example, funds made available for SLDS through the America Recovery and Reinvestment Act were tied to a requirement for the inclusion of student and teacher identifier systems, and related data on enrollment, demographics, program participation, test records, transcript information, college readiness test scores, the successful transition to postsecondary programs, enrollment in postsecondary remedial courses, entries and exits from various levels of the education system. This expansion in student PII heightens the need for comprehensive privacy and data protection plans to support the development and use of SLDS. This paper uses the tenets of Fair Information Practices and the concepts essential to data

stewardship to outline the steps required to develop a solid privacy and data protection plan or use in developing, maintaining, protecting, and using student record data.

Statewide Education Longitudinal Data Systems: A Federal/State Partnership to Support Data-Driven Decisionmaking

✦Jeff Sellers, Consultant, , *sellers.jeff@comcast.net*

Key Words: Data, data use, slds, longitudinal data, education data

Since 2005, the US Education Department has been offering competitive funds to states for the development of statewide longitudinal data systems (SLDS). More than 40 states have received SLDS grant funds to develop these systems. Other states are using Race to the Top (RT3) funds or their own state funds to build these systems, which link education information, longitudinally, to enable a robust method for tracking student performance. States are also taking their K-12 SLDS' and linking them to early learning, post-secondary, employment and other data sources in an effort to get a more comprehensive picture of students. States need to leverage the potential that comes with these SLDS' and take these opportunities to use their SLDS for state level policy and program evaluation, along with providing education data down to the districts, schools and teachers to inform instruction and facilitate data-driven decisions. This session will provide examples how Florida has been able to use its SLDS, the PK-20 Education Data Warehouse, to evaluate and build upon existing policies and programs within the state, support legislation development and future plans for additional data use.

Statewide Education Longitudinal Data Systems: A Federal/State Partnership to Support Data-Driven Decisionmaking

✦Kathy Gosa, Kansas State Department of Education, 120 SE 10th Avenue, Topeka, KS 66612, *kgosa@ksde.org*

Key Words: business intelligence, data quality, statewide longitudinal data system, enterprise data warehouse, data governance, sustainability

Kansas began development of its Statewide Longitudinal Data System in 2005 with the collection of individual student data via KIDS (Kansas Individual Data on Students). Through both federal and state funding the system has been expanded to an Enterprise Data System which includes an enterprise data warehouse and business intelligence solutions spanning early childhood though postsecondary education, and is now being connected to workforce data. With this ever-growing source of rich longitudinal data come concerns and considerations for privacy, data quality, and effective use of the data to improve education. This session will describe KSDE's journey in designing and developing this system, including the considerations for governance, data quality, sustainability, data use, and technological expansion.

281 Detection of clusters ●

IMS, ENAR, International Chinese Statistical Association, International Indian Statistical Association

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Cluster Detection Using Percolation

✦Ery Arias-Castro, Department of Mathematics, UCSD, Department of Mathematics, UCSD, 9500 Gilman Drive # 0112, La Jolla, CA , *eariasca@math.ucsd.edu*; Geoffrey R. Grimmett, University of Cambridge

Key Words: cluster detection, surveillance, scan statistic, percolation theory, multiple hypothesis testing

Consider the task of detecting a salient cluster in a sensor network, which we model as an undirected graph with a random variable attached to each node. †Motivated by recent research in environmental statistics and the drive to compete with the reigning scan statistic, we explore alternative methods based on the percolative properties of the network. †The first method is based on the size of the largest connected component after removing the nodes in the network whose value is lower than a given threshold. The second one is the upper level set scan test introduced by Patil and Taillie (2003), which consists in scanning the connected components after thresholding. †We establish their performance in an asymptotic decision theoretic framework where the network size increases to infinity, both in the context of parametric and nonparametric classes of clusters. †Percolation theory is at the base of our theoretical results, which are complemented by some numerical experiments.

Fast Multivariate Subset Scanning for Scalable Cluster Detection

✦Daniel Bertrand Neill, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, *neill@cs.cmu.edu*; Edward McFowland III , Carnegie Mellon University; Skyler Speakman, Carnegie Mellon University

Key Words: event detection, spatial scan statistics, linear-time subset scanning

We present new, fast algorithms for multivariate event detection in massive space-time datasets. We first review the linear-time subset scanning (LTSS) property, which allows efficient optimization of a likelihood ratio scan statistic over all subsets of the data. This work extends the LTSS framework from univariate to multivariate data, enabling computationally efficient detection of irregularly shaped space-time clusters even when the numbers of spatial locations and monitored data streams are large. We demonstrate that two variants of the multivariate space-time scan statistic can each be efficiently optimized over proximity-constrained subsets of locations and over all subsets of the monitored data streams, enabling timely detection and accurate characterization of emerging events. Using our fast algorithms, we compare these two multivariate scan statistics on real-world disease surveillance tasks, demonstrating tradeoffs between detection and characterization performance. Finally, we discuss extensions of LTSS to other data types, including graph and tensor data. This work was partially supported by National Science Foundation grants IIS-0916345, IIS-0911032, and IIS-0953330.

Scan Statistics for Detecting Genome Structural Variants by Paired-End Sequencing

✦Nancy Zhang, Stanford University, , *nzhang@stanford.edu*

Genomic structural variants include inversions, transpositions, and gains and deletions of large stretches of DNA. Paired-end sequencing experiments allow the detection of these variants at an unprecedented resolution. I will describe probabilistic models for this type of data, and scan statistics designed to capture the signals. Efficient algorithms and false discovery rate control for these scan statistics will also be discussed.

282 Quality Issues in Healthcare ■●

Section on Quality and Productivity, International Indian Statistical Association, SSC, WNAR, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Process Improvement in Health Care: Overall Resource Efficiency

Jeroen de Mast, IBIS UvA; ✦Benjamin Kemper, IBIS UvA, Plantage Muidergracht 12, Amsterdam, International 1018 TV The Netherlands, *b.p.h.kemper@uva.nl*

Key Words: resource management, six sigma, lean thinking, business process re-engineering, line balancing

This paper aims to develop a unifying and quantitative conceptual framework for healthcare processes from the viewpoint of operations management. The work adapts standard models from operations management to the specifics of healthcare processes. We propose concepts for organizational modeling of healthcare processes. In addition, we propose an axiological model which breaks down general performance goals into process metrics. The connection between both types of models is made explicit as a system of metrics for process flow and resource efficiency. The framework is generic, and will need modifications and refinements for specific applications. The conceptual models offer exemplars for practical support in process improvement efforts, suggesting to project leaders how to make a diagrammatic representation of a process, which data to gather, and how to analyze and diagnose a process's flow and resource utilization. By providing conceptual models and practical templates for process diagnosis, the framework relates many disconnected strands of research and application in process improvement in healthcare to the unifying pursuit of process improvement.

Quality Issues In Healthcare

✦Jason Gillikin, Spectrum Health, 100 Michigan St NE MC 157, Grand Rapids, MI 49503, *jason.gillikin@spectrum-health.org*

Key Words: hospital, quality, productivity, auditing, ROI

As part of the Quality Issues in Healthcare session organized by the Section on Quality and Productivity, this discussion -- with a less techncial approach -- will explore the various ways that hospitals can adequately calculate the productivity and accuracy of registration and coding staff, and how these measures ought to link to big-picture metrics like denied claims and claim-edit rejections. We will explore the approaches developed by one Midwestern hospital system to quantify and improve performance when no clear metrics or best practices exist industry-wide.

Advancing Systems Engineering In Healthcare

✦Victoria Jordan, MD Anderson Cancer Center, Quality Engineering and Clinical Operations Informatics, , *vsjordan@mdanderson.org*

Abstract: In this presentation, the systems engineering approach to reforming and improving various facets of the healthcare system will be discussed. The University of Texas healthcare enterprise is actively implementing tools such as process optimization, heuristics, and simulation throughout their healthcare system. Some examples of the implementation of these quality tools and results will be provided.

Healthcare Quality Engineering: Current Practices And Needs

✦James Benneyan, Northeastern University, Center for Health Organization Transformation, Boston, 02115, *benneyan@coe.neu.edu*

Problems with our healthcare system are well-known and staggering, including poor access, inefficient processes, equity disparities, practice variability, and patient safety issues, all at enormous costs. In response and among other approaches, significant efforts are focusing on industrial front-line process improvement methods. These include six sigma, PDSA, lean, and other front-line approaches that all rely on significant testing and interpreting data - although typically with very little rigor in even basic statistical methods to support this work. This talk is divided roughly into thirds - discussing the current use of standard statistical quality engineering methods within this context, several disturbing trends and pitfalls, and potential opportunities and research needs to help accelerate improvement of healthcare processes.

283 Outstanding Innovations in Statistics Education: Past, Present and a Glimpse at the Future ●

ASA/NCTM Joint Committee on Curriculum in Probability and Statistics, Section on Statistical Education

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Outstanding Innovations in Statistics Education: Past, Present, and a Glimpse at the Future

✦Amy G Froelich, Iowa State University, Department of Statistics, 3109 Snedecor Hall, Ames, IA 50011-1210 USA, *amyf@iastate.edu*; ✦Roger Woodard, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695-8203, *roger_woodard@ncsu.edu*; ✦Jo Hardin, Pomona College, 610 N. College Ave., Millikan 226D, Claremont, CA 91711, *jo.hardin@pomona.edu*

Key Words: statistics education, quantitative literacy

This panel session will be comprised of three recent recipients of the Waller Education Award from the Section on Statistical Education and the American Statistical Association, given for innovation in the instruction of elementary statistics. Each panelist will briefly present his/her academic background and path to getting involved in statistics education followed by a longer discussion of areas of his/her previous and current work that has impacted the broader statistics education movement. Areas of discussion will include: developing a new under-

graduate course for a broad audience emphasizing quantitative literacy, implementing new curricula in statistics for current and future secondary mathematics teachers, and developing new course materials for teaching statistical inference in the introductory statistics course. Each panelist will then offer his/her ideas of important areas for future innovation in statistics education.

284 Controversies in the philosophy of Bayesian statistics

General Methodology, International Chinese Statistical Association, International Indian Statistical Association, Section on Bayesian Statistical Science

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Controversies in the Philosophy of Bayesian Statistics

✦James Berger, Duke University, , *berger@stat.duke.edu*; ✦Christian P Robert, Universite Paris-Dauphine, CEREMADE, Paris cedex 16, International 75775 France, *xian@ceremade.dauphine.fr*; ✦Cosma Shalizi, Carnegie Mellon University, , *cosma.shalizi@gmail.com*; ✦Robert Kass, Carnegie Mellon University, , *kass@stat.cmu.edu*

Key Words: Bayes, philosophy, subjective probability

Debates about the philosophy of statistics can affect statistical practice. Bayesians and non-Bayesians give different recommendations in areas ranging from experimental design to interval estimation to causal inference. Within Bayesian inference, subjective and objective priors can yield very different posterior inferences. The speakers in this panel bring a range of perspectives regarding questions of inductive and deductive inference, subjective and objective probability, and classical and Bayesian modes of inference. We hope with this panel to relate some of these philosophical differences to statistical theory and practice.

285 New multi-scale and connectivity methods for brain imaging data ■

Biometrics Section, ENAR

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Connectivity And Causality In Brain Imaging

✦Martin A Lindquist, Columbia University, 1255 Amsterdam Ave, Room 1031, 10th Floor, MC 4690, New York, NY 10027, *martin@stat.columbia.edu*

Key Words: brain, fMRI, connectivity, causal inference, structural equation models

To date human brain mapping has primarily been used to construct maps indicating regions of the brain that are activated by certain tasks. Recently, there has been an increased interest in augmenting this type of analysis with connectivity studies that seek to describe how brain regions interact and how these interactions depend on experimental conditions and behavioral measures. Often researchers discriminate between functional connectivity, the undirected association between two or more fMRI time series, and effective connectivity, the directed influence of one brain region on the physiological activity recorded in other brain regions. In this talk we argue that this distinction is not entirely clear or relevant. Instead, the validity of the conclusions made from any connectivity method will depend strongly on certain key assumptions which are often poorly specified and difficult to check. We conclude by showing how ideas from causal inference can provide a mathematical framework for determining these assumptions.

Multiscale Adaptive Spatial-Temporal Models for Functional Image Data

✦Hongtu Zhu, University of North Carolina Department of Biostatistics, 3101 McGavran-Greenberg, CB#7420, Chapel Hill, NC 27599 USA, *hzhu@bios.unc.edu*; Jianqing Fan, Princeton University; Weili Lin, ENAR, IMS, ASA; japing Wang, ENAR, IMS, ASA

Key Words: multiscale adaptive, spatial-temporal, functional imaging, wavelet thresholding

We develop a multiscale adaptive spatial-temporal model (MASTM) for functional imaging data with complex spatial patterns on a two-dimensional (2D) slice or in a 3D volume. Most statistical methods, including wavelet thresholding, focus on independently estimating functional activity at each spatial location (voxel). MASTM, however, is to specifically count for the complex spatial patterns in functional images and build a comprehensive statistical model in order to simultaneously estimate all unknown functions across all voxels. MASTM has four features: being spatial, being connected, being hierarchical, and being adaptive. MASTM analyzes all observations in the ellipsoid of each voxel and its homogeneous groups. These consecutively connected ellipsoids across all voxels can capture spatial dependence among imaging observations, while these homogeneous groups allow for the combination of spatially disconnected clusters. Our simulation studies and real data analysis confirm that MASTM significantly outperforms the existing methods. This is joint work with Japing Wang, Weili Lin, and Jianqing Fan.

Generalized Spectral Measures of Cross-Dependence

✦Hernando Ombao, Brown University, 121 South Main Street, 7th Floor, Providence, RI 02912 USA, *ombao@stat.brown.edu*

Key Words: Multivariate time series, Electroencephalograms, Coherence, Spectral Analysis, Fourier analysis

Coherence is one common measure of cross-dependence between components in multivariate time series. Under the classical Cram\'er representation of stochastic processes, cross-coherence at a single frequency ω is the squared magnitude of the cross-correlation between the random increments at frequency ω of two time series.It identifies frequency bands that drive linear association between signals. However, classical coherence may not fully capture the dependence in complex signals such as electroencephalograms (EEGs). In this talk, we extend the concept of coherence at a single frequency to coherence at dual frequency (a pair of frequencies) at both contemporaneous and lagged time blocks. Under this novel concept, one may investigate how oscillatory activity at frequency ω at the (b-1)-th time block predict activity at frequency λ at the b-th time block. We develop simple estimators for dual coherence based on replicated trials

and derive their asymptotic distributions. Our results generalize the classical results on coherence analysis. These novel measures will be utilized to analyze EEG data recorded during a visual-motor experiment.

Twinmarm: Two-Stage Multiscale Adaptive Regression Methods For Twin Neuroimaging Data

✦Yimei Li, St. Jude Children's Research Hospital, 262 Danny Thomas PL, Memphis, TN 38103, *Yimei.Li@stjude.org*; Hongtu Zhu, University of North Carolina Department of Biostatistics; japing Wang, ENAR, IMS, ASA

Key Words: Twin, imaging, multi-scale

Twin imaging studies have been valuable for understanding relative contribution of environment and genes on brain structure and function. Conventional analyses of twin imaging data include three sequential steps: spatially smoothing imaging data, independently fitting a structural equation model at each voxel, and finally correcting for multiple comparisons. However, conventional analyses suffer from the same amount of smoothing throughout the whole image, the arbitrary choice of smoothing extent, and the decreased power in detecting environmental and genetic effects introduced by smoothing raw images. The goal of this article is to develop a two-stage multiscale adaptive regression method (TwinMARM) for spatial and adaptive analysis of twin neuroimaging and behavioral data. TwinMARM consists of two stages and each stage is a spatial, hierarchical, and adaptive procedure. The first stage is to establish the relationship between twin imaging data and a set of covariates of interest, such as age and gender. The second stage is to distangle the environmental and genetic influences on brain structure and function.

286 Statistical Modeling in Service Science ■●

Business and Economic Statistics Section

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

The Impact Of Individual Decisions On The Equity Of H1N1 Vaccine Distribution

✦Jessica L. Heier Stamm, Kansas State University, 2023 Durland Hall, Manhattan, KS 66506, *jlhs@k-state.edu*; Nicoleta Serban, Georgia Institute of Technology; Julie Swann, Georgia Institute of Technology

Key Words: spatial statistics, equity, access, public health logistics, decentralized systems

We integrate spatial statistics, integer programming, game theory, and geographical information systems to provide insights for improved resource allocation in systems where individual choices impact system outcomes. We apply our models to shipment location and vaccine quantity data from the early stage of 2009-2010 H1N1 vaccination campaign, during which vaccine demand significantly outpaced supply. Using integer programming and game theory, we model individuals' choices among vaccination sites. The model output is a measure of vaccine accessibility at the census tract level in terms of distance traveled and number of people per vaccine (scarcity) at each site. We contrast the results of this model, which incorporates individual choice, with a traditional optimization model that assumes a centralized planner controls all choices to optimize access. We then introduce spatial statistical models to identify systematic correlations between access to vaccination sites and socioeconomic factors across geographic space. The results point to geographic inequity in vaccine accessibility. The differences are more pronounced in the model that explicitly captures individual choice.

Theory and Methodology for Exoneration Data

✦Kobi Ako Abayomi, Georgia Tech, 765 Ferst Dr., 444 Groseclose, Atlanta, GA 30332, *kobi@gatech.edu*; Jessica Gabel, Georgia State University; Otis Brian Jennings, Duke University

Key Words: Discrete Multivariate Distributions, Statistical Dependence, Statistics in Law, Statistical Classification, Discrete Dependence, Case-Control Methods

In 2009, the work of Innocence Network member organizations led to the exoneration of 27 people in the United States. Since 1992, the Innocence Projects have helped over 250 wrongly convicted persons prove their innocence and gain freedom. Unfortunately, these exonerations are likely a miniscule sample of the number of wrongly convicted persons. Recent research suggests that the number innocents languishing in prison may be greater than 28,500, in non-death penalty cases alone. Many of these cases slip through the cracks: DNA evidence is unavailable, non-existent or insufficient to merit review for exoneration. We suggest methodology for the possible determinants of exoneration using a case-control setup for discrete multivariate dependent data; the raw data the Innocence Projects gather - court records, criminal histories, personal and local demographics. We illustrate the method using a sample from the Georgia and North Carolina Innocence Projects.

A Space-Time Varying Coefficient Model: The Equity of Service Accessibility

✦Nicoleta Serban, Georgia Institute of Technology, 765 Ferst Dr, Atlanta, GA 30331, *nserban@isye.gatech.edu*

Key Words: service accessibility, spatial-temporal modeling, varying coefficient model, service distribution equity

Research in examining the equity of service accessibility has emerged as economic and social equity advocates recognized that where people live influences their opportunities for economic development, access to quality healthcare, and political participation. In this research, service accessibility equity is concerned with where and when services have been and are accessed by different groups of people, identified by location or underlying socioeconomic variables. Exploring distance-based accessibility measures and using new statistical methods for modeling spatial-temporal data, this paper estimates demographic association patterns to financial service accessibility varying over a large geographic area (Georgia) and over a period of 13 years. The underlying model is a space-time varying coefficient model including both separable space and time varying coefficients and space-time interaction terms.

Data-Driven Workforce Management In Labor-Intensive Service Sytems

✦Haipeng Shen, University of North Carolina at Chapel HIll, 353 Hanes Hall, Chapel Hill, NC 27599 US, *shenhaipeng@gmail.com*

Key Words: call center, health care, arrival rate undertainty, agent heterogeneity, forecasting, queueing model

Labor-intensive service operations, such as Telephone Call Centers or Emergency Departments in hospitals, are traditionally analyzed as queueing systems using mathematical queueing models. Recently, statisticians started to supplement these mathematical models with theoretically-interesting and practically-relevant statistical analysis. This is enabled by the availability of transaction-level (or call-by-call) data bases. Empirical analysis of such data has validated in some cases, and refuted in others, the applicability of existing queueing models to such operations. This has also stimulated the development of further models that capture previously unaccounted-for phenomena, such as arrival-rate uncertainty and server heterogeneity. I shall present some ongoing research aiming at addressing such phenomena.

A Statistical Approach To Forecast Multi-Step Process Durations

✦Alejandro Veen, IBM Research, , *av7000@gmail.com*

Key Words: time-to-event data, duration forecasting, non-parametric, survival analysis

The management of multi-step processes is an important problem faced by a large number of businesses. A wide variety of day-to-day operations can in fact be described as multi-step business processes in fields ranging from transportation and manufacturing to various application processes such as applications for licenses, patents, loans, insurance coverage, and even employment positions. One aspect of particular interest is forecasting the duration of the process for a unit of interest (e.g. the application for a government-issued license), given its current step in the process pipeline and additional factors. Instead of explicitly modeling the queuing dynamics, this talk presents a flexible statistical approach based on non-parametric survival analysis. After presenting an example, the strengths and limitations of this approach will be discussed for different types of service science problems.

287 Exploration of Data Quality in Work with Administrative Records and Sample Surveys ■●

Section on Government Statistics, Section on Survey Research Methods, Social Statistics Section, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Data Linkages: Self-Reported Earnings Versus Reconciliation Process For Enforcement Operations To Address Improper Payments

✦Ryan J. Machtmes, Social Security Administration, 9th Floor ITC Building, 500 E Street SW, Washington, DC 20254, *Ryan.Machtmes@ssa.gov*; Aneer Rukh-Kamaa, Social Security Administration; Renee Ferguson, Social Security Administration

Key Words: Data linkages, Administrative data, Data quality issues, Logistic regression

Improper payment of benefits resulting from inaccurate self-reports of individual earnings is one problem presently facing Federal entitlement agencies and scrutinized by oversight panels, watchdog groups, and the taxpayer. A Reconciliation process involving the linkage of Social Security administrative data with IRS earnings, rather than relying upon beneficiaries to self-report earnings from substantial gainful activity, may help reduce the expense of improper payments through early detection vis a vis this linkage. This paper explores the use of a Reconciliation process and data linkage to detect Social Security beneficiaries whom received improper payments, and will discuss the development of a logistic regression model to aid this detection. Presented also are performance metrics and some initial findings from a pilot study to indicate the effectiveness of this approach. This paper also highlights certain issues related to the data quality of self-reported earnings.

Exploration Of Data Quality In Work With Administrative Records And Sample Surveys

✦daniell toth, Bureau of Labor Statistics, , *toth.daniell@bls.gov*; Polly Phipps, Bureau of Labor Statistics

Key Words: Recursive partitioning, non-ignorable nonresponse, propensity model, establishment survey, Classification and Regression Trees (CART)

To gain insight into how characteristics of an establishment affect nonresponse, a recursive partitioning algorithm is applied to the Occupational Employment Statistics May 2006 survey data to build a regression tree. The tree models an establishment's propensity to respond to the survey given certain establishment characteristics. It provides mutually exclusive cells based on the characteristics with homogeneous response propensities. This makes it easy to identify interpretable associations between the characteristic variables and an establishment's propensity to respond; something not easily done using a logistic regression propensity model. A linear representation of the tree model is used to test the model obtained using the May data against data from the November 2006 Occupational Employment Statistics survey. This test, on a disjoint set of establishment data, gives compelling evidence that the tree model accurately estimates the response rate of establishments. This representation is then used along with frame-level administrative wage data linked to sample data to investigate the possibility of nonresponse bias. We show that there is a risk that the nonresp

Matching Sipp And Administrative Record Data

✦James A Sears, SSA, ITC Building, 9th Floor, 500 E St., SW, Washington, DC 20254, *jim.sears@ssa.gov*

Key Words: SIPP, Social Security, matching

Due to declining willingness of survey respondents to report their social security numbers, the most recent Survey of Income and Program Participation data are matched to Social Security Administration records with a statistical process that does not involve SSNs. This leads to a higher overall match rate than was achieved for earlier SIPP panels, but it could still result in a matched sample that is unrepresentative of the overall population. For example, unmatched cases may now be heavily concentrated among low-income families that do not file

income tax returns. This paper considers whether studies of matched data from the 2008 SIPP panel are likely to yield biased results, and it explores whether weight adjustments are likely to be sufficient to overcome any potential biases.

Re-Weighting The National Health And Nutrition Examination Survey Linked To Medicare Administrative Records

✦Lisa Beth Mirel, CDC/NCHS, 3311 Toledo Road, Hyattsville, MD 20782, *LMireL@cdc.gov*; Jennifer D Parker, National Center for Health Statistics

Key Words: National Health and Nutrition Examination Survey, NHANES, Centers for Medicare and Medicaid Services, CMS, Statistical weights, Non-response adjustment

The National Health and Nutrition Examination Survey (NHANES) was recently linked to Centers for Medicare and Medicaid Services (CMS) Medicare data. Non-response bias from refusal or insufficient information for linkage needs to be considered when analyzing these data. We compared statistical weight adjustments for non-response based on respondents who could be linked and had data on the 2007 Medicare Denominator file from the 1999-2004 NHANES. Different weights were calculated using SUDAAN's WTADJUST procedure. Summary statistics for a handful of health and health care measures estimated from the different weights were compared for adults 65 years of age or older. Initial results, found that estimated proportions of obese adults and estimated proportions enrolled in Medicare Advantage plans did not vary among weighting approaches. However the comparison of the different weighting approaches may differ by analysis variables, and respondent characteristics. Ongoing work, examining outcomes with different distributional properties, stratified by important demographic covariates, such as gender and race/ethnicity is in progress.

Responsive Designs For Rare Subpopulations Subject To Misclassification

✦Randall Powers, Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Washington, DC 20212, *powers_r@bls.gov*; John L. Eltinge, U.S. Bureau of Labor Statistics

Key Words: Efficient design, Measurement error, Point estimation, Response error, Subsampling, Total statistical risk

When working with data from surveys or administrative records, one often encounters special issues in the estimation of prevalence rates and means for rare subpopulations. Two notable issues arise when some data are subject to misclassification, and when one obtains limited data for some of the subpopulations of interest. This paper explores the extent to which one may adapt responsive-design methods to produce improved estimators of both prevalence rates and subpopulation means. In this paper we propose an adaptive design and methods for estimating prevalence rates and subpopulation means under this design. We then provide a detailed simulation study focused on evaluation of bias and mean squared error for the proposed estimators under specified sets of conditions.

288 New Tools for the Analysis of Large-Scale, Complex Datasets ■

Section on Statistical Computing, International Chinese Statistical Association, Section for Statistical Programmers and Analysts, SSC

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

On Mathematics Of Data

✦Yuan Yao, Peking University, School of Mathematical Sciences, 5 Yiheyuan Rd, Beijing, 100871 P.R. China, *yuany@math.pku.edu.cn*

Key Words: data analysis, algebraic topology, discrete geometry, statistical computing

Over the last two decades, the world has witnessed an enormous growth in data sets that are complex, high-dimensional, and massive. This is in part an inevitable consequence of technological advancement. Among other factors, more sophisticated instruments and sensoring devices (from gene sequencers to camera phones), new human activities in the web-enabled world have led to ever-more complicated data being created and collected on a never-before scale. Traditional techniques for analyzing data have become inadequate and it calls upon a wider scope of collaborations among computer scientists, statisticians and mathematicians, the former two groups have fostered the growth of computational statistics. In this talk, we will discuss how mathematics created for largely intellectual reasons (algebraic topology, differential geometry, harmonic analysis, etc) or for completely different purposes (e.g. processing communication signals or understanding human intelligence), could nonetheless provide powerful new tools for analyzing these modern data sets. In particular we will focus on some novel schemes arising from traditional topology and geometry in modern data analysis.

Algorithms For Machine Learning On Massive Datasets

✦Alexander Gray, Georgia Institute of Technology, Klaus Advanced Computing Building, 266 Ferst Drive, Atlanta, GA 30332, *agray@cc.gatech.edu*

Key Words: massive data, algorithms, computational methods, machine learning

I'll describe algorithms and data structures for allowing the most powerful machine learning methods, which often scale quadratically or even cubically with the number of data points, to be performed many orders of magnitude faster than naive implementations. Such techniques can make previously impossible statistical analyses tractable on the scale of entire sky surveys. I will touch on scalable algorithms we have developed for nearest-neighbors, kernel density estimation, nonparametric Bayes classification, principal component analysis, local linear regression, hidden Markov models, k-means, manifold learning, support vector machines, and n-point correlation functions, among others. In addition to techniques inspired by computational geometry, fast multipole methods, and Monte Carlo integration, we employ a distributed framework which can be thought of as a higher-order version of Google's MapReduce. Our algorithms have enabled several first-of-a-kind large-scale analyses of astronomy data, networking data, biomedical data, and others.

Statistical Learning In The Cloud With Graphlab

✦Carlos Guestrin, Carnegie Mellon University, Pittsburgh, PA 15213, *guestrin@cs.cmu.edu*

Key Words: machine learning, statistical learning, parallel algorithms, distributed algorithms, cloud computing, large-scale data

Exponentially increasing dataset sizes have driven Statistical Learning experts to explore parallel and distributed computing. Furthermore, cloud computing resources such as Amazon EC2 have become available, providing cheap and scalable platforms for large scale computation. However, due to the complexities involved in distributed design, it can be difficult for researchers to take full advantage of cloud resources. Existing high-level parallel abstractions like MapReduce are insufficiently expressive while low-level tools like MPI and Pthreads leave learning experts repeatedly solving the same design challenges. Targeting common patterns in learning , we developed GraphLab, which compactly expresses asynchronous iterative algorithms with sparse computational dependencies, while ensuring data consistency and achieving a high degree of parallel performance. We demonstrate the expressiveness of the framework by designing and implementing parallel versions for a variety of real-world tasks, including learning graphical models with approximate inference, Gibbs sampling, tensor factorization, Co-EM, Lasso and Compressed Sensing, evaluating on clouds of up to 256 processors.

Selection Of Causal Rare Variants In Sequencing Studies

✦Lin Li, Harvard University, 655 Huntington Ave, SPH2 Rm 451, Boston, MA 02115, *linli@hsph.harvard.edu*; Xihong Lin, Harvard School of Public Health

Key Words: variable selection, sequencing data, penalized regression

There has been increasing interests in studying rare variants and their role underlying human complex diseases, as they may contribute to the genetic component in disease susceptibility that is unexplained by common variants. The advances of re-sequencing methods have made such studies possible, and efforts are taken in searching for regions enriched of causal variants, both rare and common. An important step that follows is to identify individual causal variants from these regions. Naturally variable selection can be applied, but it is challenging as causal rare variants tend to be under-powered to be selected. We propose a weighted penalized regression method for variable selection favoring rare variants. The method is applied to both continuous and binary traits, and various weighting schemes are evaluated. Simulations show that our weighted method is more powerful than unweighted ones in identifying rare causal variants, while taking into account common ones. A real dataset is also studied using our method.

Large Scale Kernel Belief Propagation For Nonparametric Graphical Model

✦Le Song, Carnegie Mellon University, Lane Center and Machine Learning Department School of Computer Science, 5000 Forbes Avenue, Pittsburgh, 15217, *lesong@cs.cmu.edu*

Belief propagation is an inference algorithm for graphical models that has been widely and successfully applied in a great variety of domains. We propose a nonparametric generalization of belief propagation, Kernel Belief Propagation (KBP), for pairwise Markov random fields: messages are represented as functions in a reproducing kernel Hilbert space (RKHS), and message updates are simple linear operations in the RKHS. KBP makes none of the assumptions commonly required in classical BP algorithms: the variables need not arise from a finite domain or a Gaussian distribution, nor must their relations take any particular parametric form. Rather, the relations between variables are represented implicitly, and are learned nonparametrically from training data. KBP has the advantage that it may be used on any domain where kernels are defined (R^d, strings, groups), even where explicit parametric models are not known, or closed form expressions for the BP updates do not exist. The computational cost of message updates in KBP is polynomial in the training data size. We also propose a constant time approximate message update procedure by representing messages using a small number of basis functions. We experiment with a parallel implementation of KBP for image denoising, image to depth prediction, and protein structure prediction problem: KBP is faster than competing classical and nonparametric approaches (by orders of magnitude, in some cases), while providing significantly more accurate results.

289 Perspectives on Reproducible Research: Moving from Buzzword to Doctrine ■

Section on Statistical Consulting, Section on Statistics in Epidemiology, Committee of Representatives to AAAS, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Efficient And Effective Analysis In The Drug Approval Setting

✦William K. Mountford, PPD, 929 North Front Street, Wilmington, NC 28401-3331 USA, *kenne.mountford@ppdi.com*

Key Words: reproducible research, consulting statistics, good practice, clinical trial

Statistical reporting is usually the last step encountered in a research study. Study results and their interpretation drive decision making and cannot afford human error. Therefore, statistical results are often required to be reproducible by numerous individuals in numerous settings. For instance, clinical trial results may be calculated by numerous statisticians within a clinical research organization, pharmaceutical company, and regulatory agency. The present discussion will provide an overview of processes implemented in the clinical research organization industry, which allow for an effective and systematic approach to reproducing study results. The process begins at the data collection stage where explicit written specifications are developed to instruct someone how to map a datapoint from a clinical database into an analysis database. Details will be provided on developing an analysis plan and study result mockups as well as the result validation process. In addition, there will be a discussion of techniques that increase efficiencies of the overall process described. In closing the real world pros and cons of the described process will be discussed.

Good Practices And Research Reproducibility In The Department Of Veterans Affairs (Va) Cooperative Studies Program

✦Rebecca B McNeil, Durham VA Medical Center, Epidemiologic Research & Information Center (152), 508 Fulton Street, Durham, NC 27705, *rebecca.mcneil@va.gov*; Domenic J Reda, Hines VA Hospital

Key Words: reproducible research, consulting statistics, good practice

The VA Cooperative Studies Program (CSP) provides coordinating centers for the development and implementation of qualified investigator-initiated clinical trials. Global SOPs for CSP address issues of reproducibility and good practice using four approaches. First, the coordinating center maintains independence from the principal investigator with respect to quality management, despite acting as scientific collaborators. Second, the statistician's role is expanded to include that of team lead, with responsibility for protocol development and management of the internal merit review submission, study start-up, and extensive monitoring. Reflecting this increased level of responsibility, the statistician is included in the Planning and Executive Committees. Third, commercial co-sponsor involvement in data monitoring committee organization and manuscript preparation is prohibited. Finally, a rigorous validation process is required during the development of data management systems and statistical analysis code. We provide a detailed overview of these reproducibility and good practice approaches, and discuss their perceived effects on the statistician, team dynamics, and research quality.

Reproducible Research In An Academic Environment

✦John Kloke, University of Pittsburgh, , *klokejd@upmc.edu*

Key Words: reproducible research, consulting statistics, good practice

To implement reproducible research in an academic environment one faces several unique challenges. For example, the size of the projects vary from the analysis of data from a pilot study to the data collection, monitoring, and analysis of an R01 clinical trial. Given that we work closely with the Department of Biostatistics we have several graduate students at any one time who do may do a large portion of the analysis which results in a high degree of turn-over. We also like to provide our faculty a degree of academic freedom, perhaps less present in other areas of medical research, which introduces additional challenges. In this talk we give a brief overview of the Data Center: discuss personnel, computational resources, collaborators. Then we further discuss the challenges unique of the implementation of reproducible research systems in an academic environment. We then provide specific examples of the use of reproducible research tools.

Reproducible Research In Daily Practice

✦JoAnn Alvarez, Vanderbilt University School of Medicine, Department of Biostatistics, 1161 21st Ave South, S2323 MCN, Nashville, TN 37232, *joann.alvarez@vanderbilt.edu*; Matthew Shotwell, Vanderbilt University School of Medicine

Key Words: reproducible research, good practice, consulting statistics

Reproducible methods can reduce errors and improve the quality of research. However, concerns persist regarding the practicality of adopting these methods on a routine basis or institution-wide. The Department of Biostatistics at Vanderbilt University has 27 master's level staff biostatisticians working on collaborative research with clinicians. The great majority of them use R, Sweave, and Latex every day to prepare data, run analyses, and make graphics, tables, and written reports. The department promotes the use of reproducible methods by providing computer support staff that are knowledgeable in these methods, developing a wiki code repository, offering opportunities for continuing education in R, Sweave, and Latex at weekly R/Sweave clinics, peer continuing education talks, and monthly computing talks. This presentation addresses how statisticians can personally implement reproducible methods by explaining the workflow involved, offering solutions for situations when a final report is requested in word-processor formats, and showing examples of reports. Finally, we describe how institutions can provide support for these practices.

290 Social Statistics Section Student Paper Competition Winners

Social Statistics Section, Section on Survey Research Methods, Section on Government Statistics

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Bayesian Analysis Of Between-Group Differences In Variance Components In Hierarchical Generalized Linear Models

✦Brady Thomas West, Michigan Program in Survey Methodology, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48106-1248, *bwest@umich.edu*

Key Words: Bayesian Analysis, Variance Components, Generalized Linear Mixed Models, Likelihood Ratio Testing, Survey Interviewing, Interviewer Variance

Frequentist approaches to making inferences about the variances of random cluster effects in hierarchical generalized linear models (HGLMs) for non-normal variables have several limitations. These include reliance on asymptotic theory, questionable properties of classical likelihood ratio tests when pseudo-likelihood methods are used for estimation, and a failure to account for uncertainty in the estimation of features of prior distributions for model parameters. This paper compares and contrasts alternative approaches to making a specific type of inference about the variance components in an HGLM, focusing on the difference in variance components between two independent groups of clusters. A Bayesian approach to making inferences about these types of differences is proposed that circumvents many of the problems associated with alternative frequentist approaches. The Bayesian approach and alternative frequentist approaches are applied to an analysis of real survey data collected by independent groups of interviewers in the Continuous National Survey of Family Growth (NSFG).

Combining Information From Multiple Complex Surveys

✦Qi Dong, Program in Survey Methodology, University of Michigan, 426 Thompson Street, Room 4050, Ann Arbor, MI 48104, *qidong@umich.edu*

Key Words: complex sample survey, combining rule for multiple surveys, synthetic populations, multiple imputation, health insurance coverage rates, BRFSS, NHIS, MEPS

Increasingly many substantive research questions require a web of information that is not adequately collected in a single survey. Fortunately, survey organizations often repeatedly draw samples from the same population for different surveys and collect a considerable amount of overlapped variables. This paper proposes a principled method for combining multiple complex surveys from a missing data perspective. The basic proposal is to simulate multiple copies of the population for each survey and stack them across surveys so we can impute the missing variables by borrowing information from other surveys. Then the estimates for the statistic of interest are calculated from each complete synthetic population and are combined using the method in this paper. The more surveys we combine, the more information we have and thus the more accurate and precise the combined estimator is. The proposed method is used to combine the National Health Interview Survey and the Medical Expenditure Panel Survey to estimate health insurance coverage rates for the whole and subdomain populations. The combined estimator is more accurate and precise compared to the one from individual surveys.

Variance Inflation Factors In The Analysis Of Complex Survey Data

✦Dan Liao, RTI International, 3404 Tulane Dr, Apt 22, Hyattsville, MD 20783, *dliao@rti.org*

Key Words: cluster sample, collinearity diagnostics, linearization variance estimator, survey-weighted least squares, stratified sample, unequal weighting

Survey data are often used to fit linear regression models. The values of covariates used in modeling are not controlled as they might be in an experiment. Thus, collinearity among the covariates is an inevitable problem in the analysis of survey data. Although many books and articles have described the collinearity problem and proposed strategies to understand, assess and handle its presence, the survey literature has not provided appropriate diagnostic tools to evaluate its impact on regression estimation when the survey complexities are considered. We have developed variance inflation factors (VIFs) that measure the amount that variances of parameter estimators are increased due to having non-orthogonal predictors. The VIFs are appropriate for survey-weighted regression estimators and account for complex design features, e.g. weights, clusters, and strata. Illustrations of these methods are given using data from a household survey of health and nutrition.

Synthetic Data Generation For Small Area Estimation In The American Community Survey

✦Joseph Sakshaug, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48104 USA, *joesaks@umich.edu*

Key Words: synthetic data, small area estimation, disclosure, microdata

Demand for small area estimates is growing among a variety of stakeholders who use these data to study issues affecting local communities. Statistical agencies regularly collect data from small geographic areas but are prevented from releasing small area identifiers due to disclosure concerns. Several disclosure control methods are used to disseminate microdata, including summary tables, suppression of geographical details, and Research Data Centers, but none of these methods is ideal for meeting the growing demand for small area datasets. This research tests a new method for disseminating public-use microdata that contains more geographical details than are currently being released. Specifically, the method replaces the observed microdata with fully-synthetic, or imputed, microdata generated from a posterior predictive distribution. A hierarchical Bayesian model is used to preserve the small area inferences and simulate the synthetic data. Confidentiality protection is enhanced because no actual values are released. The synthetic data is evaluated by comparing inferences obtained from the synthetic data with observed data from the 2005-2007 American Community Survey.

Imputation And Estimation Under Nonignorable Nonresponse For Household Surveys With Missing Covariate Information

✦Anna Sikov, Hebrew University of Jerusalem, Israel, Department of Statistics, Hebrew University of Jerusalem, Har Hatsofim, Jerusalem, 91905 Israel, *anasikov@mscc.huji.ac.il*; Danny Pfeffermann, University of Southampton and Hebrew University

Key Words: Bootstrap, Calibration, Horvitz-Thompson type estimator, Respondents distribution, Nonrespondents distribution

In this research we develop and apply new methods for handling not missing at random (NMAR)nonresponse. We assume a model for the outcome variable under complete response and a model for the response probability, which is allowed to depend on the outcome and auxiliary variables. The two models define the model holding for the outcomes observed for the responding units. Our methods utilize information on the population totals of some or all of the auxiliary variables in the two models, but we do not require that the auxiliary variables are observed for the nonresponding units. We develop an algorithm for estimating the parameters governing the two models and show how to estimate the distributions of the missing covariates and outcomes. We investigate conditions for the convergence of the algorithm for parameter estimation and develop conditions for the consistency and asymptotic normality of the estimators obtained by the application of this algorithm. We also consider several test statistics for testing the model fitted to the observed data and study their performance. The new developments are illustrated using real data set collected as part of the Household Expenditure Survey.

291 Model Selection and Inference in Complex Problems with Survey Applications ■●

Section on Survey Research Methods, International Chinese Statistical Association, Social Statistics Section, SSC

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Benchmarking The Mixed Linear Model For Longitudinal Data

✦Thuan Nguyen, Oregon Health and Science University, Dept. of Public Health & Preventive Medicine, Oregon Health and Science University, Portland, OR 97239-2966 USA, *nguythua@ohsu.edu*

Key Words: Benchmarking, longitudinal data, mixed models, missing values

We propose a simple approach to missing values in longitudinal studies incorporating the linear mixed models. In many applications of the linear mixed models, there are either missing values that include the responses or the covariates or both. This is particularly the case in longitudinal data. As a result, the mixed model relation between the response and covariates is satisfied only for the complete data. We propose to supplement the complete-data mixed model relation by a system of benchmark equations that involve both complete and incomplete data, and thus make more efficient use of the available information. We set up the framework of our approach through conditional and marginal models, and study the finite sample performance of our benchmarked maximum likelihood estimators through empirical studies and compare them with the MLE based only on the complete-data.

Gee Analysis Of Clustered Binary Data With Diverging Number Of Parameters

✦Lan Wang, University of Minnesota, School of Statistics, 313 Ford Hall, 224 Church Street S.E., Minneapolis, MN 55455, *wangx346@umn.edu*; Jianhui Zhou, University of Virginia; Annie Qu, University of Illinois at Urbana-Champaign

Key Words: GEE, clustered binary data, high-dimensional data, correlated data

Clustered binary data with a large number of covariates have become increasingly common in many scientific disciplines. We consider a generalized estimating equations (GEE) approach to analyzing such data when the number of covariates grows to infinity with the number of clusters. This approach only requires the specification of the first two marginal moment conditions. The likelihood function does not need to be specified or approximated. In the first part of the talk, we consider an extension of the classical theory of GEE to the large n, diverging p framework. We provide appropriate regularity conditions and establish the asymptotic properties of the GEE estimator. In particular, we show that the GEE estimator remains consistent and asymptotically normal, and that the large sample Wald test remains valid even when the working correlation matrix is misspecified. In the second part of the talk, we propose penalized GEE for simultaneous variable selection and estimation. The properties of the penalized GEE are investigated in the ``large n, diverging p" setting which allows the possibility of p>n. . Furthermore, we propose an effective iterative algorithm to solve the penalized GEE

Threshold Estimation Using P-Values

✦Atul Mallik, University of Michigan, 439 W Hall, 1085 S University Ave, Ann Arbor, MI 48109, *atulm@umich.edu*; Moulinath Banerjee, University of Michigan; Bodhisattva Sen, Columbia University

Key Words: baseline value, change point, stump function

We seek to identify the threshold value at which a real valued function takes off from its baseline level, under regression and multiple dose-response setting. This is relevant to a broad range of problems, e.g., estimating the minimum effective dose level in certain dose-response models in pharmacology, detecting tidal disruptions in dwarf spheroidal galaxies, advent of global warming etc. An important case involves the baseline set having the form [0, d], the unknown d being the threshold. On this set, the function stays at its baseline value (minima or maxima) and then takes off. The approach involves fitting stumps to p-values obtained from tests conducted at different points/bins under the hypothesis that the function is at its baseline level. This works well owing to the fact that the p-values exhibit a dichotomous behavior. This problem has natural connections to change point estimation. The procedure is consistent under minimal conditions, involves at most one tuning parameter and is computationally easy to implement. It also attains the optimal rate of convergence under certain assumptions. The asymptotic distribution are derived and subsampling is also shown to work.

Bayesian Variable Selection For Hierarchical Spatial Regression Models

✦Chae Young Lim, Michigan State University, A 426 Wells Hall Dept of Statistics and Probability, Michigan State University, East Lansing, MI 48823 USA, *lim@stt.msu.edu*; Taps Maiti, Michigan State University; Sarat C Dass, Michigan State University

Key Words: spatial regression model, bayesian variable selection, hierarchical model

Hierarchical spatial regression models are commonly used in modeling disease incidence/mortality data. We are interested in selecting a subset of spatially varying regression coefficients which helps to identify different sets of covariates over different regions that affect disease rates. Bayseian approach is natural to handle this overparametrization. We investigate various possible priors that could leads us to select such sebsets of covariates using simulation study.

Joint Estimation Of Multiple Gaussian Graphical Models By Nonconvex Penalty Functions With An Application To Genomic Data

✦Hyonho Chun, Purdue University, , *chunh@purdue.edu*

Key Words: GGMs, regularization, gene networks

Inferring unknown gene regulation networks is one of key questions in systems bi- ology with important applications such as understanding disease physiology and drug discovery. These applications require inferring multiple networks in order to reveal the differences among different conditions. The multiple networks can be in- ferred by Gaussian graphical models by introducing sparsity on the inverse covari- ance matrices via penalization either individually or jointly. We propose a class of nonconvex penalty functions for the joint estimation of multiple Gaussian graphical models. Our approach is capable of regularizing both common and condition spe- cific associations without explicit parametrization as well as has oracle property for both common and specific associations. We show the performance of our nonconvex penalty functions by simulation study and then apply it to real genomic dataset.

292 Advanced Methods for Predicting Survival Outcomes

Biometrics Section, ENAR, International Indian Statistical Association, Section on Health Policy Statistics, Section on Risk Analysis

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Estimating Subject-Specific Treatment Differences For Risk-Benefit Assessment With Event-Time Data In The Presence Of Competing Risks

✦Brian Claggett, Harvard School of Public Health, Boston, MA , *bclagget@hsph.harvard.edu*; Lihui Zhao, Harvard School of Public Health; LJ Wei, Harvard University; Lu Tian, Stanford University; Davide Castagno, University of Turin

Key Words: Clinical trial, Nonparametric estimation, Personalized medicine, Survival analysis

To evaluate treatment efficacy using event-time data from a comparative study, one usually makes inference about a summary measure which quantifies an overall treatment difference. However, a positive result based on such a measure does not mean that every future subject should be treated by the new therapy. It is desirable to identify subjects, using baseline covariates, who would benefit from the new treatment from a risk-benefit perspective. In this paper, we propose a systematic approach to achieve this goal with competing risk event-time data. First, we utilize data from a similar, but independent, study to build a parametric score for stratifying the current study patients. We then use the present study to obtain a nonparametric estimate of the treatment difference, with respect to each event, for any fixed score. Confidence interval and band estimates are constructed to quantify uncertainty in our inferences for the treatment differences over the score. To illustrate the new proposal, we use data from two cardiovascular studies for evaluating specific beta-blockers. The score is based on time to death, and the competing events are death, MI, hospitalization and toxicity.

One Weighted Complete-Case Method For Competing Risks Model With Missing Cause Of Failure

✦Xiongwen Tang, The University of Iowa, 10 Village Dr APT 7, North Liberty, IA 52317, *xiongwen-tang@uiowa.edu*; Ying Zhang, Department of Biostatistics, The University of Iowa

Key Words: competing risks, weighted, complete-case, multiple imputation, proportional hazards

We study the proportional hazards model to analyze the competing risks data, where the cause of failure could be missing due to informative reasons. The possible causes of failure can be generalized as two types: one for causes with primary interest and one for the others. The proportional hazards model with different baseline hazards and regression coeffients is specified for each cause type respectively. A weighted complete-case (WCC) method is studied by assuming a parametric model for the missing mechanism. Asymptotic properties have been further explored. We finally compare our WCC method to the ones proposed in Goetghebeur and Ryan (1995) and Lu and Tsiatis (2001) by showing an example of breast cancer data from ECOG study E1178.

Modeling Left-Truncated And Right-Censored Survival Data With Longitudinal Covariates

✦Yu-Ru Su, University of California, Davis, One Shields Ave, Davis, CA 95616, *yrsu@ucdavis.edu*; Jane-Ling Wang, University of California, Davis

Key Words: Likelihood approach, Biased sample, EM algorithm, Monte Carlo integrations

There is a surge in medical follow-up studies to include longitudinal covariates in the modeling of survival data. To jointly model the survival time and longitudinal process under the presence of biased sample due to left-truncation, we propose a modified likelihood approach for statistical inference since the direct extension from cases subject to right-censoring to additional left-truncation is not trivial. Due to the random effects on modeling the longitudinal processes, an EM algorithm is employed to locate the nonparametric maximum modified likelihood estimate (NPMMLE). Although the sample is biased, the resulted NPMMLE of the regression coefficient in the survival component can be estimated unbiased while the coefficients in the longitudinal component cannot be recovered. The consistency and asymptotic normality of the NPMMLE of the survival-related parameters will be illustrated. To verify the performance of the proposed estimates, we conducted some simulations with different levels of variation of the measurement errors and the latent variables. The application of the proposed methods on a multi-center AIDS cohort study provides an illustration from the practical aspect.

Estimating The Survival In The Presence Of Dependent Truncation

✦Jing Qian, Harvard School of Public Health, 655 Huntington Ave, 4th floor, Boston, MA 02115, *jqian@hsph.harvard.edu*; Rebecca A. Betensky, Harvard School of Public Health, Harvard University

Key Words: Copula, Kendall's tau, Peterson bounds, Product limit estimator, Quasi-independence

An increasing number of clinical trials and observational studies are assembled using complex sampling involving truncation. Ignoring the issue of truncation or incorrectly assuming quasi-independence can lead to bias and incorrect results. Currently available approaches for dependently truncated data are sparse and incomplete. In this paper, we propose a product limit estimator for survival under dependent truncation using a hazard ratio assumption linking the unobservable region to the observable region. We also derive nonparametric sharp bounds for the survival and bounds for survival which are based on reasonable assumptions on a hazard ratio function using the proposed product limit estimator. The properties of these bounds are discussed. The proposed method is applied to some real data examples.

Quantifying The Association Between Disease Onset And Lifetime Under Cross-Sectional Sampling

✦Marco Carone, University of California, Berkeley, , *marcocarone@gmail.com*; Daniel O. Scharfstein, Johns Hopkins Bloomberg School of Public Health; Masoud Asgharian, McGill University

Key Words: cross-sectional sampling, disease onset, measures of association, semiparametric model, successive durations, survival analysis

Common measures of association usually quantify the departure of random variables from independence. Such measures often do not provide a scientifically meaningful interpretation when the variables are successive durations with a constrained sum, such as age at disease onset and time from onset to death. We propose a novel semiparametric model to quantify in a sensible manner the impact of disease on an individual's lifetime. A class of estimating equations for the parameters of this model are constructed and inference using data from a cross-sectional survey with longitudinal follow-up is discussed. Such data often arise in the study of the natural history of a disease. Because these data are subject to both systematic biases and loss to follow-up, their analysis is challenging. The asymptotic properties of the proposed inferential procedures are derived, and data from the Canadian Study of Health and Aging are analyzed to learn about the impact of dementia on the lifetime of elderly Canadians.

293 Issues in Development of Drugs and Biologies

Biopharmaceutical Section

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Parallelism For A Sigmoid Curve In Assay Development

✦Jason Liao, Teva Branded Pharmaceutical Products R&D, Inc, 425 Privet Road, Horsham, PA 190444, *jason.liao@tevausa.com*

Key Words: Bioassay, Sigmoid curve, Transformed data, Relative potency, Parallelism, Equivalence limits

For a meaningful comparison of the potency for different biological agents/products, a fundamental assumption is the curve from the test sample needs to be parallel to the curve of the standard sample. The common approach for assessing parallelism is to conduct either a significant test or an equivalent test on the parameters from the sigmoid dose response curve such as the commonly used symmetric four-parameter logistic function (4PL) or asymmetric five-parameter logistic function (5PL). However, there are many drawbacks for these types of approaches. It is very intractable to implement the equivalence test on multiple parameters of a nonlinear model because of the difficulty to obtain the joint confidence region and the complexity of the joint confidence region. To overcome these drawbacks of the current approaches, an equivalent approach directly on the response difference, rather than on the curve parameters, is proposed to assess the parallelism in this paper. The equivalence limit is established as the reference against the reference itself and the limit can be easily scientific and subject knowledge judged to see if it is practically feasible or not.

Regression To The Rescue: The Use Of Statistical Doe To Derive A Novel Manufacturing Control Strategy For Ensuring The Quality Of An Antibody-Drug Conjugate

✦Lisa J Bernstein, Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080 USA, *bernlisa.lisa@gene.com*; Daniel A Coleman, Genentech, Inc.

Key Words: antibody-drug conjugate (ADC), design of experiments (DOE), split-plot, biologic, manufacturing

ADC-X is an antibody-drug conjugate therapeutic candidate consisting of a monoclonal antibody linked to a potent cytotoxin by a chemical linker. The Drug-to-Antibody Ratio (DAR), the average number of toxin molecules per antibody, is a critical quality attribute related to both safety and efficacy; hence it must be controlled in manufacturing. This talk describes how a sequential experimentation strategy characterized the ADC-X manufacturing process. Multiple laboratory constraints on experimentation were managed by incorporating blocking and split-plotting into fractional-factorial designs. These characterization studies identified the important factors in each process step and provided a regression model relating those factors to the DAR and other attributes. The presentation will also show how the designed experiments uncovered a serious challenge to commercial ADC-X manufacturing and then provided a tool that met the challenge: the regression model formed the basis for a novel process control strategy involving an adjustable setpoint for the amount of linker in each batch. This control strategy has been successfully implemented in large-scale manufacturing.

Statistical Considerations For Defining Cut Points And Titers In Anti-Drug Antibody (Ada) Assays

✦Ken Goldberg, Johnson & Johnson, 965 Chesterbrook Blvd, Mail Stop C-4-1, Wayne, PA 19087, *kgoldber@its.jnj.com*; Sheng Dai, Johnson & Johnson; Allen Schantz, Johnson & Johnson

Key Words: Immunoassay, Anti-drug antibody, Cut point, Titer

Biologic drug products can induce immune response, which in turn can adversely affect safety or efficacy so it is important to develop reliable laboratory test methods. Two ADA assays are presented. Nonlinear regression is used in each to help choose a screening or a titer cut point. The "Screening cut point" is the response above which undiluted samples are classified as reactive. Dilution of a reactive sample above its titer drops its signal below the "Titer cut point". In an RIA to detect anti-drug antibodies, our problem is to choose an adjusted signal (%binding) definition from many functions of various assay controls. The screening cut point has 5% false positive rate by definition. We choose the %binding formula with the lowest limit of detection. In a separate titration assay, the mean response for pure assay diluent is greater than for naÔve negative samples. Since high titration samples approach pure diluent and high signal, the titer cut point must be greater than the screening cut point. A method for choosing the titer cut point is developed based on the titer's CV so that intra-assay variability adds negligibly to inherent uncertainty due to the titer's discreteness.

Statistical Issues In Estimating, Tracking, And Comparing Complex Non-Linear Parameters Determined When Developing And Manufacturing Biologics

✦Patrick J Gaffney, ImClone Systems, 33 ImClone Drive, Branchburg, NJ 08876, *Patrick.Gaffney@ImClone.com*; Anthony Lonardo, ImClone Systems

Key Words: biologics, Bayesian, cell growth

Measurement is a key concern in development and manufacture of biologics. It is important in establishing the design space and in subsequent control in the manufacturing process. This paper looks at measurement in the final bioreactor where the antibody is primarily produced. Traditionally the nutrients (e.g. glucose and glutamine), by-products (e.g. lactate and ammonia) and the bioreactors conditions (e.g. pH, dissolved oxygen and carbon dioxide) are tracked. This paper examines measurement of the cell culture growth, which ultimately captures how the cell culture responds. While the cell growth rate in earlier bioreactors is, for the most part, exponential, the growth pattern of the cell culture in this bioreactor is more complex and exhibits distinct phases of growth and death. Two methodologies are examined - one based on the Bayesian paradigm and the other using a frequentist framework. This paper will explore the unique statistical issues associated with precise estimation and control.

294 Recent Developments in Regression ■●

IMS, International Indian Statistical Association

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Bayesian Logistic Regression For Medical Claims Data Using Cpu And Gpu

✦Ivan Zorych, Columbia Univ., Room 1005 SSW, 1255 Amsterdam Ave., New York, NY 10027, *zorych@gmail.com*; Patrick Ryan, Johnson & Johnson; David Madigan, Columbia University

Key Words: Bayesian regression, coordinate descent, logistic, lasso, GPU, CUDA

Bayesian logistic regression for medical claims data is a novel statistical approach that possesses the advantages of regression analysis such as being resistant to confounding by co-medication and adjusting for masking effect. Mapping medical claims data into the form appropriate for the regression analysis is an essential step. We consider several ways to represent claims data in the form appropriate for regression. We investigate Bayesian regression models with either Normal or Laplace priors. Analysis of each condition of interest requires fitting a separate regression model. Fitting such a model is a challenging computational task because each dataset contains millions of reports and thousands of covariates. Our numerical approach to logistic regression relies on coordinate descent algorithm. We consider two implementations of this algorithm, traditional central processing unit, CPU, version and a parallel implementation that utilizes graphics processing unit, GPU. The performance of our approach will be illustrated on the simulated and real data.

The Development Of Coordinate-Descent Algorithms - A Review

✦Wenjiang Fu, Michigan State University, East Lansing, MI 48824, *fuw@msu.edu*

Key Words: Coordinate-descent, Efficiency, Lasso penalty, Least-squares regression, Regularization

Thanks to the Lasso penalty model (Tibshirani 1996), a new series of coordinate-descent algorithms have been developed to fit models for high dimensional data, including the Shooting algorithm, the least squares support vector machine algorithm, the coordinate-descent algorithms and application to high dimensional genome data, high dimensional generalized linear regression models, and the active shooting algorithm. The coordinate-descent algorithm can be traced back to the cyclic coordinate-descent algorithm described in Luenberger's monograph Linear and Nonlinear Programming (1972). These algorithms share the same feature, unlike the classical method of working with high dimensional matrix decomposition, such as the Cholesky's decomposition, they take the coordinate-wise minimization procedure iteratively to achieve efficient computation, particularly for high dimensional data, usually of 10,000 dimension or higher. The computational advantages of these methods have no doubt been appreciated. However, certain issues may still exist, including the convergence rate and the efficiency for data that possess largely different characteristics. I will discuss these issues in this talk.

Coordinate Descent Algorithms For Lasso Problems: The Good, The Bad, And The Ugly

✦Noah Simon, Stanford University, 390 Serra Mall, Stanford University, Stanford, CA 94305, *nsimon@stanford.edu*

Key Words: LASSO, coordinate descent, algorithms

Coordinate descent and block-wise descent algorithms, lauded for their simplicity and efficacy, have been used with much success to fit LASSO (and LASSO-like) problems. However there is some misunderstanding about when coordinate descent is appropriate. In this talk we examine aspects of lasso-like problems that make them particularly suited to coordinate descent and contrast these with more ill-suited problems. We look at potential difficulties with fitting by coordinate descent, and note when and how these difficulties can be overcome. Parts of this talk will be based on joint work with Rob Tibshirani, and inspired by conversations with Jerome Friedman and Trevor Hastie.

A Novel Coordinate-Descent Algorithm For Median Regression

Wenjiang Fu, Michigan State University; ✦Martina Fu, Michigan State University / Okemos HS, East Lansing, MI 48824, *fumartin@msu.edu*

Key Words: Coordinate-Descent, Lasso penalty, Median regression, Oracle properties, Robust

Compared to the least-squares regression, median regression is more robust to extreme data values. The development of computer technology facilitates the algorithms for median regression and makes it promising to fit complex data. Yet, fast and efficient algorithms are desirable for both least-squares regression and median regression, in particular, when the study data are sparse in high dimensional space. In this talk,

we will present a novel coordinate-descent algorithm for median regression and will further extend it to the penalized median regression with the Lasso (L1) penalty. The new algorithms follow the same spirit of the Shooting algorithm for the least-squares regression with the Lasso penalty, and enjoy the fashion of simple coordinate-descent iteration similar to the Shooting algorithm. We demonstrate with simulations and real data that the new algorithms make the once complex algorithms simpler and easy for programming, and efficient to implement. We also demonstrate that the Lasso penalty median regression enjoys the oracle properties as studied in Xu and Ying (2008).

Doubly Regularized Cox Regression For High-Dimensional Survival Data Via Cyclic Coordinate Descent

Tongtong Wu, University of Maryland; ✦Sijian Wang, University of Wisconsin, Madison, 600 Highland Ave., Madison, WI 53705, *swang@biostat.wisc.edu*

Key Words: Coordinate descent, Cox regression, High-dimensional, Regularization, Survival, Variable selection

In many scientific applications, there is a natural grouping of predictors. For example, in biological applications, assayed genes or proteins can be grouped by biological roles or biological pathways. Usually, people want to identify both important groups and important variables within selected groups. In this talk, we propose a doubly regularized Cox regression for survival data. Our regularized objective function in convex, and the method can achieve variable selection in both group level and individual level. We also developed a fast algorithm via cyclic coordinate descent method. Cyclic coordinate descent avoids matrix operations since parameters are updated one by one, which yields a very fast computing speed. It is also numerically stable due to the lack of matrix operations for large systems. We demonstrate our method and algorithm using both simulation studies and a real ovarian cancer dataset. This is a joint work with Tongtong Wu at University of Maryland.

295 Bayesian Modeling and Applications

Section on Bayesian Statistical Science, International Indian Statistical Association

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

On Bayesian inference and prediction for computer software

✦Nuria Torrado, Universidad Carlos III, Calle Madrid, 126, Getafe, Madrid, 28903 Spain, *nuria.torrado@uc3m.es*; Rosa E Lillo, Universidad Carlos III; Michael P Wiper, Universidad Carlos III

Key Words: Bayesian analysis, software reliability, nonhomogeneous Poisson processes

The main purposes of this talk is to describe statistical inference and prediction for software reliability models in the presence of covariate information. In particular, we develop a semi-parametric, Bayesian model to estimate the numbers of software failures over various time periods when it is assumed that the software is changed after each time period. Goodness-of-fit testing of the model are developed using a deviance information criterion, and predictive inferences on future failures are shown. Real life examples are presented to illustrate the new model.

Bayesian Dynamic Regression Models For Interval Censored Survival Data

✦Xiaojing Wang, Department of Statistics, University of Connecticut, 215 Glenbrook Rd. U-4120, Storrs, CT 06269, *xiaojing.wang@uconn.edu*; Ming-Hui Chen, University of Connecticut; Jun Yan, University of Connecticut

Key Words: Cox model, Latent variables, Markov chain Monte Carlo, Reversible jump, Semiparametric, Time-varying coefficients

The Cox model with time-varying coefficients offers great flexibility in capturing the temporal covariate effects. Methodology development for time-varying coefficient models, however, has been largely limited to right censored data, with very limited work on interval censored data frequently arising in medical and public health research, where the event time is not observed exactly but only known to be between a time interval. Further, analysts need to specify which coefficients are time-varying and which are not at the time of fitting. In a Bayesian framework, we propose a dynamic Cox regression model for interval censored data, where the coefficient curves are piece-wise constant, but the number of pieces and the jump points are covariate specific and estimated from the data. The model automatically determines the extent to which the temporal dynamics is needed for each covariate, resulting in smoother, more stable, and more efficient curve estimates. With carefully chosen prior, the posterior computation is carried out via an efficient reversible jump MCMC algorithm. Two real data applications show the competitiveness of the proposed model and reveal some new findings.

Adaptive Gaussian Predictive Process Models for Large Spatial Datasets

✦RAJARSHI GUHANIYOGI, DIVISION OF BIOSTATISTICS, UNIVERSITY OF MINNESOTA, A460 MAYO BUILDING, MAIL CODE 303, 420 DELAWARE STREET S.E., MINNEAPOLIS, MN MN55455, *guhan003@umn.edu*; Andrew Oliver Finley, Michigan State University; SUDIPTO BANERJEE, DIVISION OF BIOSTATISTICS, UNIVERSITY OF MINNESOTA; Alan E Gelfand, Department of Statistical Science

Key Words: Bayesian hierarchical models, Intensity surfaces, Low-rank models, Markov chain Monte Carlo, Poisson process, Predictive Process

Large point referenced datasets occur frequently in the environmental and natural sciences. Use of Bayesian hierarchical spatial models is undermined by the onerous computational burden caused in fitting these large spatial datasets. Low-rank spatial process models attempt to resolve this problem by projecting spatial effects to a lower-dimensional subspace. This subspace is determined by a judicious choice of ``knots'' or locations that are fixed a priori. One specific approach considers representations in terms of lower-dimensional realizations. One such representation leads to a class of predictive process models (e.g. Banerjee et al., 2008) for spatial and spatiotemporal data. Our contribution here expands upon predictive process templates that fix the locations which determine the lower-dimensional realization. We explore sto-

chastic modeling of the knots, viewing them as a point pattern. Using such adaptive specifications can reduce the number of knots required to adequately inform about the underlying process realizations, yielding substantial computational benefits.

Emulating a gravity model to infer the spatiotemporal

✦Roman Jandarov, Department of Statistics, Penn State University, 315 W Beaver Ave Apt 12, State College, PA 16801 USA, *raj153@psu.edu*; Murali Haran, Pennsylvania State University; Ottar Bjornstad, Department of Entomology and Center for Infectious Disease Dynamics, Penn State University; Bryan Grenfell, Department of Ecology and Evolutionary Biology, Princeton University

Key Words: Gravity Model, Disease Dynamics, Measles, Bayesian Inference, Gaussian Processes, SIR

We study a metapopulation model for regional measles dynamics that uses a gravity coupling model and a time series susceptible-infected-recovered (T-SIR) model for local dynamics. Standard maximum likelihood or Bayesian inference for this model is infeasible as there are potentially tens of thousands of latent variables in the model and each evaluation of the likelihood is expensive. We develop an efficient discretized MCMC algorithm for Bayesian inference with these expensive likelihood evaluations. However, we find through a simulation study that parameter estimates are biased and simulations at the obtained parameter settings do not explain some important biological characteristics of the data. We propose fitting a Gaussian process (GP) model to forward simulations of the gravity model at a number of parameter settings. Treating this GP model as an approximation ('emulator') for the gravity model, we perform a full Bayesian analysis of a given data set. This approach allows us to conveniently study posterior distributions of the key parameters of the gravity model and has number of advantages over the classic likelihood based inference.

Travel Time Estimation For Emergency Vehicles Using Bayesian Data Augmentation

✦Bradford S. Westgate, Cornell University, 113 Stewart Ave. #4, Ithaca, NY 14850, *bsw62@cornell.edu*

Key Words: Travel times, EMS, Ambulance, Reversible-jump, MCMC, Gibbs sampling

Estimates of ambulance travel times on road networks are required for effective emergency medical services (EMS) planning. We introduce a new method for estimating the distribution of travel times on each road segment in a city, using data from Global Positioning System (GPS) devices on the ambulances. Due to sparseness and error in the GPS data, the exact ambulance paths and travel times on each road segment are not known. To estimate the travel time distributions using this data, we must infer the path driven from the GPS data. This is called the "map-matching" problem. We consider the unknown paths and travel times to be missing data, and simultaneously estimate them and the road segment travel time parameters using Bayesian data augmentation. We evaluate our method as a solution to the map-matching problem and the travel time distribution problem on a small region of Toronto, using a simulation study and data from Toronto EMS. We compare mean travel time estimates from our method with estimates from a speed averaging method, representative of currently used techniques.

296 Beyond Intro Statistics: Additional Topics to Excite and Lead Students Further

Section on Statistical Education

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Beyond Intro Statistics: Additional Topics To Excite And Lead Students Further

✦James Bush, Waynesburg University, 51 W. College Street, Waynesburg, PA 15370 USA, *jbush@waynesburg.edu*; ✦Michael A Costello, Bethesda-Chevy Chase High School, American University, 3901 CATHEDRAL AVE NW APT 616, 616, Washington, DC 20016 US, *MichaelAVCostello@gmail.com*; ✦Ann Cannon, Cornell College, 600 First Street SW,, Mount Vernon, IA 52314, *ACannon@cornellcollege.edu*

Key Words: education, introductory, teach, curriculum, regression, college

How can professors and teachers supplement the introductory statistics curriculum with topics that broaden our students' view of the field and entice them to further their study of statistics? What should we be teaching students interested in taking a second course in applied statistics? What can teachers of AP Statistics do to get students excited about a career in statistics (or a statistics-related field)? This panel will include practical examples of advanced topics in statistics in the context of other disciplines, field trip and panel ideas, and curriculum information for a second course in statistics.

297 Missing Data and Measurement Error

Biometrics Section, Biopharmaceutical Section, ENAR, International Chinese Statistical Association, Section on Health Policy Statistics, International Indian Statistical Association, Section on Statistics in Epidemiology, Section on Quality and Productivity, Section on Survey Research Methods

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Multiple Imputation of High-Dimensional Mixed Missing Data

✦Ren He, University of California Los Angeles, 51-254 CHS Building, UCLA, Los Angeles, CA 90095, *renhe@ucla.edu*; Thomas R Belin, UCLA Dept. of Biostatistics

Key Words: Multiple Imputation, Hierarchical prior, MCMC, Parameter-Extended algorithm

It is common in applied research to have large numbers of variables with mixed data types (continuous, binary, ordinal or nomial) measures on a modest number of cases. Also, even a simple imputation model can be overparameterized when the number of variables is moderately large. Finding a joint model to accommodate multivariate data with mixed

data types is challenging. Here, I develop a joint multiple imputation model with multivariate normal components for continuous variables and latent-normal components for categorical variables. Following the strategy of Boscardin and Weiss (2003) and using Parameter-expanded Metropolis-Hastings estimation (Boscardin,Zhang and Belin 2008), I use a hierarchical prior for the covariance matrix centered around a parametric family. This not only substantially reduces the dimension of the parameter space but also allows the data to depart from a tightly structured covariance matrix. The statistical properties of the method are illustrated in several simulation settings.

Adjusting For Non-Response In Population-Based Case Control Studies

✦Alastair Scott, University of Auckland, Auckland, New Zealand, *a.scott@auckland.ac.nz*; Chris J. Wild, University of Auckland

Key Words: Stratified case-control studies, Non-response, Conditional maximum likelihood, Weighted estimation, Semiparametric methods

In this paper we discuss the analysis of data from population-based case-control studies when there is appreciable non-response. We develop a class of estimating equations that adjust for the non-response and yet are relatively easy to implement. For some important special cases, we also provide efficient semiparametric maximum-likelihood methods. We compare the methods in a simulation study based on data from the Women's Cardiovascular Health Study discussed in Arbogast et al (2002). Reference. Arbogast, P.G., Lin, D.Y., Siscovick, D.S., and Schwartz, S.M. (2002). Estimating incidence rates from population-based case-control studies in the presence of nonrespondents. Biometrical Journal, 44, 227--239

A Measurement Error Model For Self-Reported Physical Activity

✦Janet A. Tooze, Wake Forest University Health Sciences, Department of Biostatistical Sciences, Medical Center Boulevard, Winston-Salem, NC 27157, *jtooze@wfubmc.edu*; Richard P. Troiano, National Cancer Institute; Raymond James Carroll, Texas A&M University; Laurence S Freedman, Gertner Institute for Epidemiology and Health Policy Research

Key Words: Measurement error, Questionnaires, Validation, Physical Activity Assessment

Self-reported physical activity assessment instruments are prone to error yet systematic investigations into the structure of this measurement error are lacking. We propose a measurement error model for physical activity assessment instruments using physical activity level (PAL), the ratio of total energy expenditure (TEE) to basal energy expenditure (BEE). A physical activity questionnaire (PAQ) was administered to 451 participants aged 40-70 y in the Observing Protein and Energy Nutrition (OPEN) Study. MET minutes from the PAQ were used to estimate PAL. The main objective is to relate the PAL measurement from the PAQ to true PAL. Although true PAL was not observed, TEE was measured using an unbiased biomarker conforming to a classical measurement error model and BEE was estimated from an equation. Because BEE is estimated from an equation, it is prone to Berkson error. Therefore, the non-questionnaire measure of PAL has a mix of classical (TEE) and Berkson error (BEE). We present a measurement error model for PAL that accommodates this mixture of errors and use it to establish the relationship between the PAQ measure of PAL and true PAL and its application from the OPEN study.

A Semi-Parametric Roc Approach To Assessing Biomarkers Subject To A Measurement Error And Limit Of Detection

✦Weijie Chen, Food and Drug Administration, , *weijie.chen@fda.hhs.gov*

Key Words: ROC analysis, biomarker, measurement error, limit of detection, maximum likelihood estimation, semi-parametric model

Quantitative biomarkers are emerging to discriminate between two clinically useful conditions. The measured biomarker levels are often corrupted with a random error, which leads to an ROC curve lower than that of the true levels of biomarkers. A solution for correcting such random errors is to repeat the measurements. Due to the limit of detection (LoD) of the instruments, the biomarkers are deemed to be immeasurable when the measured level is below some threshold. Parametric ROC methods have been proposed to analyze repeated measurements of biomarkers. However, parametric methods rely on a strong assumption that the data follows a normal distribution. We investigated a semi-parametric ROC approach that relies on a much looser assumption: the data are related to normal distributions by an implicit monotonic transformation. Maximum likelihood estimation is used to estimate the error-corrected ROC parameters and quantify the amount of measurement error. Extensive simulations show that our method is robust across a broad spectrum of experimental conditions including large measurement errors, substantial LoD, and deviations from the normal distribution.

Nonparametric Estimation Of A Heaping Mechanism From Precise And Heaped Self-Report Data

✦Sandra D. Griffith, University of Pennsylvania, 503 Blockley Hall, 423 Guardian Dr, Philadelphia, PA 19104, *sgrif@upenn.edu*; Saul Shiffman, University of Pittsburgh; Daniel Heitjan, University of Pennsylvania

Key Words: measurement error, smoking cessation, digit preference, heaping

Self-report data commonly exhibit heaping: a form of measurement error that occurs when quantities are reported with varying levels of precision. Digit preference is a special case of heaping where the preferred values are round numbers. Daily cigarette counts, for example, commonly exhibit heaps at multiples of 2, 5, 10, and 20 when measured by retrospective recall. As heaping can introduce substantial bias to estimates, conclusions drawn from data subject to heaping are suspect. In the absence of more precise measurements, methods to estimate the true underlying distribution from heaped data depend on unverifiable assumptions about the heaping mechanism. A data set in which subjects reported cigarette consumption by both a precise method (ecological momentary assessment using a hand-held electronic device) and a retrospective recall method (timeline followback) allows us to forgo the usual assumptions. To exploit these unique data, we propose a nonparametric method to estimate the conditional distribution of

the heaping mechanism given the precise measurement. Application to our data suggests that recall errors are a more important source of bias than actual heaping.

Hot-Deck Multiple Imputation Via Predictive Moment Matching

✦Chia-Ning Wang, Dept. of Biostatistics, University of Michigan, , *cnwang@umich.edu*; Rod Little, University of Michigan

Key Words: Missing data, imputation, Hot-deck

Imputations, an method for handling missing data, are draws from a predictive distribution of the missing values, estimated from the observed data. The Hot-Deck method creates the distribution from "similar" responding units. In predictive mean matching, the similarity is measured by the closeness of predicted means of the incomplete variable regressed on the observed variables. Since this approach selects the matching sets on the predicted means, other key features of the predictive distribution such as variances are not taken into account. This may lead to bias when data are heteroscedastic, and there are covariates related to variance but not mean. We propose a generalization of predictive mean matching, Predictive Moment Matching, where the matching set is selected based on the predicted mean and variance simultaneously, which ensures better predictive distributions and imputations.

Maximum-Likelihood-Based Multiple Imputation

✦Tejas A Desai, Adani Institute of Infrastructure Management, 25 Saurashtra Society, Paldi, Ahmedabad, International 380007 India, *tejasdesai4@gmail.com*

Key Words: Imputation, Fisher Information, Missing data, maximum likelihood, Frequentist Analysis, General Location Model

Donald Rubin pioneered the use of Bayesian multiple imputation for analyzing a wide variety of incomplete data. Specifically, the general location model was proposed and used to impute entire data sets. Desai and Sen (2006, 2008) developed a frequentist method for analyzing randomly incomplete data without imputation by characterizing the underlying Fisher information appropriately. However, there are situations where imputation is necessary. In this paper, we propose and demonstrate the use of maximum-likelihood-based multiple imputation. After briefly outlining the theory, we present simulations and an example.

298 Regularization and Model Selection

Section on Bayesian Statistical Science, International Indian Statistical Association, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Bayesian Regularization Via The Graph Laplacian Prior

✦Sounak Chakraborty, university of missouri-columbia, 209F Middlebush Hall, Columbia, MO 65211, *chakrabortys@missouri.edu*; Fei Liu, IBM Watson Research Center; Fan Li, Duke University; Yan Liu, University of Southern California, Los Angles

Key Words: Bayesian, Elastic Net, Graph Laplacian, Grouping, Ridge regression, Lasso

Regularization is an important approach to prevent overfitting in regression. Under the Bayesian paradigm, many regularization techniques correspond to imposing certain shrinkage prior distributions on the regression coefficients. Existing Bayesian methods usually assume independence between explanatory variables a priori. In this article, we propose a novel Bayesian approach, which explicitly models the dependence structure between variables through a graph Laplacian matrix. We generalize the graph Laplacian to allow both positive and negative correlations. A prior distribution for the graph Laplacian is then proposed, which allows conjugacy and thereby greatly simplifies the computation. We show that the proposed Bayesian model leads to proper posterior distribution. Connection is made between the proposed method and some existing regularization approaches, such as the Lasso, the Elastic Net, and ridge regression. An efficient MCMC method based on parameter augmentation is developed for posterior computation. Finally, we demonstrate the method through simulation studies and a real data analysis.

Bayesian Tests On Components Of The Compound Symmetry Covariance Matrix

✦Joris Mulder, Tilburg University, Tilburg, International 5000LE the Netherlands, *j.mulder3@uvt.nl*; Jean-Paul Fox, University of Twente

Key Words: Bayes factor, compound symmetry, covariance matrices, Gibbs Sampler, intra-class correlation

Complex dependency structures are often conditionally modeled, where random effects parameters are used to specify the natural heterogeneity in the population. When interest is focused on the dependency structure, inferences can be made from a complex covariance matrix using a marginal modeling approach. In this marginal modeling framework, assumptions about conditional independence and random effects distributions are not required. Furthermore, testing covariance parameters is not a boundary problem in the marginal framework. In this paper, Bayesian tests on covariance parameter(s) of the compound symmetry structure are proposed assuming multivariate normally distributed observations. Innovative proper prior distributions are introduced for the covariance components such that the positive definiteness of the (compound symmetry) covariance matrix is ensured. The proposed priors on the covariance parameters lead to balanced Bayes factors for testing inequality constrained hypotheses. As an illustration, the proposed Bayes factors are used for testing (non-)invariant intraclass correlations across public and Catholic schools using the 1982 High School and Beyond survey data.

Bayesian Model Selection In Spatial Lattice Models

✦Joon Jin Song, University of Arkansas, AR 72701, *jjsong@uark.edu*; Victor De Oliveira , University of Texas at San Antonio

Key Words: Bayes factors, CAR models, SAR models, Jeffreys prior, Spatial data, Weight matrix

This work describes a Bayesian approach for model selection in Gaussian conditional autoregressive models and Gaussian simultaneous autoregressive models that are commonly used to describe spatial lattice data. The approach is aimed at situations when all competing models have the same mean structure, and the model differences rely on some aspects of the covariance structure. As the selection criterion the method uses posterior model probabilities computed using some default priors on the model parameters. The proposed method is illustrated using two real datasets.

The Multiple Bayesian Elastic Net

✦Hongxia Yang, Statistical Analysis & Forecasting Group, Mathematical Sciences Department,IBM, 7 Lake Street, Apt 4-O, White Plains, NY 10603, *yangho@us.ibm.com*; David Banks, Duke University; Juan Vivar, Department of Statistical Science, Duke University; David Dunson, Duke University

Key Words: Multi-task learning, Correlated predictors, Variable selection, Multiple shrinkage, Penalized estimation, Mixture Prior

We propose the multiple Bayesian elastic net (abbreviated as MBEN), a new regularization and variable selection method. High dimensional and highly correlated data are commonplace. In such situations, maximum likelihood procedures typically fail-their estimates are unstable, and have large variance. To address this problem, a number of shrinkage methods have been proposed, including ridge regression, the lasso and the elastic net; these methods encourage coefficients to be near zero (in fact, the lasso and elastic net perform variable selection by forcing some regression coefficients to equal zero). In this paper we describe a semiparametric approach that allows shrinkage to multiple locations, where the location and scale parameters are assigned Dirichlet process hyper-priors. The MBEN prior encourages variables to cluster, so that strongly correlated predictors tend to be in or out of the model together. We apply the MBEN prior to a multi-task learning (MTL) problem, using text data from the Wikipedia. An efficient MCMC algorithm and an automated Monte Carlo EM algorithm enable fast computation in high dimensions.

Robust Matrix Shrinkage Priors With Applications To Multivariate Regression

✦Minghui Shi, Department of Statistical Science, Duke University, 214 Old Chem, Box 90251, Duke University, Durham, NC 27708, *ms193@stat.duke.edu*; David Dunson, Duke University

Key Words: High-dimensional, Matrix normal, Variable selection, Regularization, Shrinkage, Sparsity

There is a rich literature on shrinkage and variable selection methods for high-dimensional regression and classification with vector-valued parameters, with the relevance vector machine (RVM) providing one widely-used method. In multivariate regression, one instead has matrix-valued parameters, with rows corresponding to responses and columns to covariates. In inducing sparsity in such matrices, it is appealing to maintain the matrix structure, allowing dependence in shrinkage within rows and columns. To address this, we propose a scale mixture of matrix normal priors to induce a Bayesian robust shrinkage model (BRSM) for multivariate regression. Our proposed model obtains robustness of handling both unknown sparsity and large outlying signals. A Gibbs sampler and a fast algorithm are developed for efficient learning of the parameters in the model, and our BRSM is shown to yield excellent performance in both simulated examples and a real application.

A New Prior For The Unconditioned Covariance Matrix

✦Samprit Banerjee, Weill Cornell Medical College, New York, NY 10065, *sab2028@med.cornell.edu*; Stefano Monni, Weill Cornell Medical College

Key Words: covariance matrix, reference prior, high dimensional, hit and run

Estimation of the covariance matrix, especially in higher dimensions ("large p small n") is a challenging statistical problem which is of great interest in many applications. It is well known that the sample covariance matrix is a poor estimator even for moderately high p. The currently accepted best estimator for the unconditioned covariance matrix is that based on the reference prior. We propose a new prior (reference-like) and demonstrate the improved estimation for higher dimensional matrices via simulations. We provide a Markov Chain Monte Carlo algorithm to implement the computation and highlight key aspects required to sample efficiently.

Efficient Factor-Analytic Priors For Correlation Matrices

✦Jared Murray, Duke University Dept of Statistical Science, , *jmurray.1022@gmail.com*; Lawrence Carin, Duke University; David Dunson, Duke University ; Joe Lucas, Duke Institute for Genome Sciences and Policy

Key Words: Bayesian, Factor analysis, Correlation matrix, Parameter expansion, Data augmentation

We introduce a new class of computationally efficient priors for correlation matrices via parameter expansion and data augmentation. Using a factor-analytic representation of the correlation matrix we are able to avoid expensive matrix inversions during MCMC sampling. In contrast with some other parameter-expanded priors our induced prior on the correlation matrix is of known form and readily analyzed, allowing for informative specifications. This prior not only regularizes estimators of the correlation matrix but also provides a decomposition analogous to traditional factor analysis and model-based principal component analysis which is of inferential and exploratory interest on its own.

299 Variable Selection and Testing

Biometrics Section, International Indian Statistical Association

Tuesday, August 2, 8:30 a.m.–10:20 p.m.

The Group Mcp For Hierarchical Variable Selection In High-Dimensional Logistic Regression

✦Dingfeng Jiang, Department of Biostatistics, The University of Iowa, C22 General Hospital, 200 Hawkins Drive, Iowa city, 52242, *dingfeng-jiang@uiowa.edu*; Jian Huang, University of Iowa; Ying Zhang, Department of Biostatistics, The University of Iowa

Key Words: variable selection, group structure, 1-norm group MCP, 2-norm group MCP, high dimensional logistic regression

Hierarchical structure exists naturally in many variable selection problems. Several methods such as group Lasso (Yuan and Lin 2006; Meier et al 2008) and group bridge (Huang et al 2009; Breheny and Huang 2009) have been proposed to account for group information. This paper proposes the composite of the minimax concave penalty (MCP, Zhang 2010) and the L1 or L2 norm of the coefficients for grouped variables for variable selection. The 1-norm group MCP enables bi-level selection at group and individual levels, while the 2-norm group MCP performs selection at group level. Under such setting, the group lasso can be viewed as a special case of the 2-norm group MCP. Simulation results show that for high-dimensional logistic models with grouping structure, both grouped MCPs have better predictive power than the ungrouped MCP. The grouped MCPs have similar false discover rate and group false discover rate, and both outperform the ungrouped MCP. The grouped MCPs tend to favor models with fewer groups. The 2-norm group MCP is sparser than the 1-norm one at group level. The application of the proposed methods is demonstrated on three microarray gene expression datasets.

Forward Stagewise Shrinkage And Addition For High And Ultrahigh Dimensional Censored Regression

✦Zifang Guo, North Carolina State University, Department of Statistics, North Carolina State University, Campus Box 8203, Raleigh, NC 27695, *zguo2@ncsu.edu*; Wenbin Lu, North Carolina State University; Lexin Li, Department of Statistics. North Carolina State University

Key Words: Adaptive LASSO, boosting, forward stagewise regression, proportional hazards model, variable selection

Despite the thriving development of variable selection methods in recent years, modeling and selection of high and ultrahigh dimensional censored regression remain challenging. When the number of predictors p far exceeds the number of observational units n, computations of many methods become difficult or even infeasible. Censoring of the outcome variable adds further complications. In this article, we propose a forward stagewise shrinkage and addition method for simultaneous model estimation and variable selection in Cox proportional hazards models with high and ultrahigh dimensional covariates. Our proposal extends a popular statistical learning technique, the boosting method, by explicitly performing variable selection and substantially reducing the number of iterations for algorithm convergence. It also inherits the flexible nature of the boosting and is straightforward to extend to nonlinear Cox models. Our intensive numerical analyses demonstrate that the new method enjoys an equally competitive performance as the best players of the existing solutions in Cox models with $p < n$, whereas it achieves a considerably superior performance than the alternative solutions when $p > n$.

Surrogate Variable Analysis Using Partial Least Squares In Gene Expression Studies

✦SUTIRTHA CHAKRABORTY, UNIVERSITY OF LOUISVILLE, 786 RAYMOND KENT COURT, ROOM NO - 3, LOUISVILLE, KY 40217, *sutirtha_sutir@yahoo.co.in*; Somath Datta, UNIVERSITY OF LOUISVILLE; Susmita Datta, UNIVERSITY OF LOUISVILLE

Key Words: Microarray, Expression heterogeneity, Confounders, Partial least squares

A primary objective in gene expression studies is to identify the genes that are differentially expressed between two different types of tissue samples. In a standard analysis, an ANOVA model is used for estimating the gene-tissue type interaction effects and to identify the differentially expressed genes. In this talk we are concerned with hidden confounders in gene expression arrays that are latent variables attributable to different biological, environmental or other relevant factors that are not accounted for. These factors may introduce spurious signals of expression heterogeneity that may lead to erroneous conclusions from such analyses. These distortions, if different for different genes, cannot be removed by simple array normalization. In this work we have explored the use of partial least squares to identify the hidden effects of the underlying latent factors which are then included in a linear model for gene expression. We have tested our methods on both simulated as well as real life gene expression data sets and have found that it detects the differentially expressed genes with much higher sensitivity and specificity compared to the standard analysis.

Assessing The Significance Of A Gene Set

✦Huey-Miin Hsueh, Department of Statistics, National Chengchi University, 64, Sec. 2, Zhinan Rd., Taipei, International 11605 Taiwan, *hsueh@nccu.edu.tw*; Chen-An Tsai, Graduate Institute of Biostatistics & Biostatistics Center. China Medical University ; Da-Wei Zhuo, Department of Statistics, National Chengchi University

Key Words: Gene set analysis, permutation, p-value, random forest

In DNA microarray studies, a gene-set analysis (GSA) is used to evaluate the association between the expression of biological pathways, or a priori defined gene sets, and a particular phenotype. Two types of differentially expressed testing are of research interest: the competitive testing and the self-contained testing. The competitive test is to determine whether the specific gene set is relatively differentially expressed when compared to other gene sets. The self-contained test is interested in finding whether the gene set alone is differentially expressed. The two tests involve different null distributions. To take consideration on the interaction or correlation within the gene set, we consider assessing the significance of the gene set by the performance of a classifier developed upon the gene set. In this study, the Random Forest classification is applied. For each of the two tests, the corresponding empirical P-value of an observed out-of-bag (OOB) error rate of the classifier is introduced by using adequate resampling method. Several real examples are analyzed for comparison. A simulation study is conducted for verification.

A Restricted Empirical Bayes Approach To Detecting Genetic Association

✦Zhenyu Yang, University of Ottawa, 451 Smyth Road, Ottawa, ON K1H 8M5 Canada, *zyang009@uottawa.ca*; David R. Bickel, University of Ottawa

Key Words: multiple comparison procedure, multiple testing, empirical Bayes, local false discovery rate, GWAS, constrained likelihood

In microarray data analysis, measurements of the expression of thousands of genes enable simultaneously testing thousands of null hypotheses, where each null hypothesis says a particular gene is not differentially expressed across the conditions studied. Similarly, in a genome-wide association study, measurements of the genotypes across hundreds of thousands of DNA sites enable simultaneously testing hundreds of thousands of null hypotheses, where each null hypothesis says a particular site is not associated with a trait. At the hypothesis testing level, the main difference between GWAS data and microarray data is not the number of hypotheses tested but rather is the fact that p, the proportion of false null hypotheses, is much smaller in the case of genetic association data. (Leading geneticists put p for GWASs within an order of magnitude of 1 in 100,000.) We find that the standard false discovery rate (FDR) methods tend to identify many times as many sites as associated with disease than is biologically plausible in light of the smallness of p. By construction, a recent restricted-parameter method of estimating the local FDR (arXiv:1104.0341) yields much more tenable results.

Universal Dependency Prediction and Variable Selection with the Mira

✦Hesen Peng, Emory University, 1518 Clifton Rd 3F, Atlanta, GA 30329, *hesen.peng@emory.edu*; Tianwei Yu , Emory University

Key Words: high-dimensional data, universal dependency, nonlinear, variable selection, prediction, Mira

The emergence of high-throughput data requires machine learning methods that accommodates universal types of dependency of arbitrary dimension. In this paper we propose the Mira score, a novel measure capable of identifying the existence of all types of probabilistic dependency (linear and nonlinear) of any dimension. Pre-Mira, a computationally efficient variable selection and prediction procedure is also proposed. Comparison and connection with existing method will also be provided.

300 Bayesian Methodology and Applications in Social Sciences - 2

Section on Bayesian Statistical Science, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Bayesian Finite Mixture Factor Analyzer With Mixed Continuous And Ordinal Responses

✦Xinming An , Department of Psychology, UCLA, 1285 Franz Hall, Box 951563, Los Angeles, CA 90095-1563, *xinming.an@gmail.com*; Peter M Bentler, UCLA

Key Words: factor analysis, finite mixture, mixed responses

Researches in the social and biomedical sciences often encounter high dimensional responses from a heterogeneous population. Finite mixture factor analyzer provides an efficient modeling technique to explore the unobserved group structures of high dimensional data. Because of its remarkable modeling ability, finite mixture factor analyzer has been widely used in many fields, such as image analysis, cluster analysis and high dimensional data visualization. However, this modeling technique cannot be applied to categorical responses. The purpose of the present research is to solve this problem by developing a Bayesian finite mixture factor analyzer with mixed continuous and ordinal responses. In our modeling process, natural conjugate priors will be used for model parameters, and the Gibbs sampler will be used to draw parameter samples from their posterior distributions. Based on these samples, point estimates and corresponding standard errors for model parameters will be developed. Finally, simulation studies will be used to investigate the properties of this modeling technique.

Hierarchical Actor-Partner Interdependence Models For Multilevel Intervention Studies

✦Li-Jung Liang, University of California, Los Angeles, 10940 Wilshire Blvd, Suite 1223, Los Angeles, CA 90024, *liangl@ucla.edu*; Li Li, University of California, Los Angeles

Key Words: Actor-Partner Interdependence Model, Multilevel Intervention, Hierarchical Longitudinal Model, HIV/AIDS

The concept of multilevel intervention is based on the assumption that individual behavior is interwoven with multiple layers; one person's predictor may influence not only that person's outcome measure, but also that person's partners' outcome measures. In this talk, I will briefly review the actor-partner interdependence models (APIMs) used in the study of family system when data from multiple family members are gathered. Next, we propose two Bayesian longitudinal APIMs, which are hierarchical longitudinal model with actor's and partner's predictors, to estimate effects of interest as well as various levels of correlations. We illustrate our approach using data from HIV-affected families in China.

A Mixture Model For The Joint Analysis Of Latent Developmental Trajectories And Survival

✦Rinke Klein Entink, TNO, PO Box 360, Zeist, 3700 AJ Netherlands, *rinke@kleinentink.eu*; Jean-Paul Fox, University of Twente; Ardo van den Hout, MRC Biostatistics Unit, Institute of Public Health

Key Words: Bayesian, MCMC, Mixture modeling, multilevel item response modeling

A joint modeling framework is proposed that integrates a mixture multilevel item response component to model latent developmental trajectories, given polytomous response data, with a survival component for continuous time survival data. The joint model is illustrated in a real data setting, where the utility of longitudinally measured cognitive function as a predictor for survival is investigated in a group of elderly persons. Time-dependent cognitive function is measured using the generalized partial credit model given occasion-specific minimental state examination (MMSE) response data. The mixture model identifies subpopulations that are relatively homogenous in their latent growth trajectories of cognitive function. A parametric survival model is stratified on these subpopulations, and cognitive function as a continuous latent variable is included as a time-varying explanatory variable, along with other covariates. Within the Bayesian framework, a Markov chain Monte Carlo algorithm is developed for simultaneous estimation of the joint model parameters. Practical issues as model building and assessment are addressed using the DIC and various posterior predictive tests.

Bayesian Models For Replicated Preference Testing

✦Suzanne Dubnicka, Kansas State University, , *dubnicka@k-state.edu*

Key Words: Bayesian modeling, choice behavior, sensory analysis, transition models

Sensory science uses various techniques to measure, analyze, and interpret human responses to products as perceived through their senses of touch, taste, sight, smell, or sound. Sensory science is often used to test consumer preferences regarding new products and to test the acceptance of new formulations of existing products. This talk focues on preference testing which is used to determine if consumers prefer one product over another. Replicated preference tests, in which consumers are asked to express their preferences for one of two products on multiple occasions, are used not only to determine consumer preference but also to evaluate changes in preference over time and strength of preference. Bayesian models are developed to address these and other questions in replicated preference testing. The advantages of using such models in sensory testing will be emphasized, and methods will be illustrated on real data.

A Predictive Likelihood Approach To Possible Endogeneity - An Application With The Us Income-Education Data: Trade-Off Between Estimation Precision And The Necessity Of Instruments

✦Nalan Basturk, Erasmus University Rotterdam, Erasmus University Rotterdam, Econometric, Instutite, Room H11-12, P.O. Box 1738, Rotterdam, International 3000 DR The Netherlands, *basturk@ese.eur.nl*; Lennart Hoogerheide, Erasmus University Rotterdam ; Herman K. van Dijk, Erasmus University Rotterdam

Key Words: predictive likelihoods, IV models, returns to education

A simple regression of earned income on years of education in order to measure the education-income effect may suffer from endogeneity problems, for example as a result of unobserved individual capabilities. The typical treatment of such issues is the use of Instrumental Variable (IV) models, where instruments are used to infer the effect of the endogenous explanatory variable. Two issues in IV models, weak instruments and endogenous instruments are addressed. We define alternative models accounting for these issues, and assess the extent to which these models are suitable to the data. For assessing model performance, we rely on Bayesian methods, as they provide general probabilistic tools to account for parameter and model uncertainty. We propose a predictive likelihood approach instead of the conventional marginal likelihood approach, where the former allows us to refrain from imposing strong prior information. We apply the proposed method to simulated data with differing degrees of endogeneity and instrument strength, and US data on the income and education. We show that this method can be used to weight the evidence of different models and to address issues in IV models.

Simplex Factor Models For Multivariate Unordered Categorical Data

✦Anirban Bhattacharya, Duke University, , *anib86@gmail.com*; David Dunson, Duke University

Key Words: Classification, Contingency table, Factor analysis, Latent variable, Nonparametric Bayes, Non-negative tensor factorization

Gaussian latent factor models are routinely used for modeling of dependence in continuous, binary and ordered categorical data. For unordered categorical variables, Gaussian latent factor models lead to challenging computation and overly complex modeling structures. As an alternative, we propose a novel class of simplex factor models. In the single factor case, the model treats the different categorical outcomes as independent with unknown marginals. The model can characterize highly flexible dependence structures parsimoniously with few factors, and as factors are added, any multivariate categorical data distribution can be accurately approximated. Using a Bayesian approach for computation and inferences, a highly efficient MCMC algorithm is proposed that scales well with increasing dimension, with the number of factors treated as unknown. We develop an efficient proposal for updating the base probability vector in hierarchical Dirichlet models. Theoretical properties are described and we evaluate the approach through simulation examples. Applications are described for modeling dependence in nucleotide sequences and prediction from high-dimensional categorical features.

Bayesian Predictive Inference From The Half-Normal Model Given A Type Ii Censored Sample

✦Hafiz M. R. Khan, Florida International University, Department of Epidemiology & Biostatistics, Robert Stempel College of Public Health, Miami, FL 33199 USA, *hmkhan@fiu.edu*

Key Words: Censored sample, Half-normal model, Likelihood function, Bayesian approach, Posterior density function, Predictive inference.

In this paper, we derive the likelihood function and the posterior density function for the parameter assuming that the given type II censored sample follows a half-normal model. By making use of the posterior density function, we derive the predictive density for a single future response, a bivariate future response, and several future responses. A Bayesian framework has been utilized in conjunction with an informative prior to derive the predictive results on the basis of a type II censored sample. Simulated type II censored samples from a half-normal model are utilized to illustrate the results.

301 Modern Approaches to Analyzing DNA Data

Biometrics Section, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Rearranging Computational Operations Greatly Improves the Accuracy of Genomic Copy Number Analysis of Tumors

✦Stan Pounds, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, *stanley.pounds@stjude.org*; Arzu Onar-Thomas, St. Jude Children's Research Hospital; Cuilan Lani Gao, St. Jude Children's Research Hospital

Key Words: genomics, cancer, change-point, microarray, sequencing, normalization

In genomic copy number analysis of tumors, a series of computational data analysis operations are performed to address a set of statistical problems. A "typical" method (1) normalizes signals of tumors and controls, (2) compares normalized signals of tumors and controls, (3) identifies copy number change-point loci, and (4) infers the copy number status of genomic segments. By systematically describing the informative biological and statistical relationships between operations, we developed a new method that (1) compares unnormalized signals of tumors and controls, (2) iteratively normalizes signal differences and infers change-point loci, (3) identifies two-copy genomic segments, and (4) infers the copy number status of genomic segments. For a leukemia data set, the new method's results show 95-99% agreement with cytogenetic validation data for tumors with simple or complex genomes. In contrast, the "typical" approach shows 85-95% and 10% agreement for tumors with simple and complex genomes, respectively. Our results show that major gains in statistical accuracy may be achieved by rearranging the "typical" order of computational operations in the analysis of genomic data.

Bioinformatics Of High-Throughput Insertional Mutagenesis To Identify Cancer Causing Genes

✦Jean A Roayaei, NIH-National Cancer Institute, 18 Miller Drive, Frederick, MD 20712, *jean_roayaei@msn.com*; Keiko Akagi, OSU Cancer Center

Key Words: CIS, Founders Mutation, Breast cancer-subtypes, Poisson Distribution, MCMC Gibbs Sampling, Human Genome

We investigated Common Insertional Sites (CISs) for human cancer genome to identify cancer causing genes in breast cancer. We considered different window sizes of 1KB, 2KB, 5KB and 10KB to identify genes that have been implicated in different breast cancer sub-types. Cancer gene mutations are rare events whether they are founder's mutations (the first incidence of breast cancer in a family) or in women who have genetic suseptibility to breast cancer. We used a Poisson probability distribution to simulate the number of cancer causing genes by comparing a random set versus a test set of different window sizes for Common Insertional Sites. We identified cancer causing genes for different breast cancer sub-types that were not implicated in previous NCI breast cancer studies. This could be used by clinicians to develop new treatments for different patients' cohorts with different breast cancer subtypes.

Quantification Of Real Time Polymerase Chain Reaction (Pcr) Data

✦Xuelin Huang, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Unit 1411, Houston, TX 77230-1402 USA, *xlhuang@mdanderson.org*; Wei Wei, MD Anderson Cancer Center; Ning Jing, The University of Texas Health Science Center at Houston

Key Words: Bioinformatics, DNA, Microarray, Polymerase chain reaction, PCR, Quantification

The quantitative real-time polymerase chain reaction (PCR) can now amplify even a tiny amount of DNA material to a detectable level and then back-calculate its initial number of molecules from the real-time PCR curve. It has been the golden standard used in modern molecular biology research to confirm discoveries by other techniques, such as microarrays and single nucleotide polymorphism (SNP) arrays. However, so far, most of the methods for the aforementioned back-calculation from PCR curves were developed by biologists. There is a great room for improvement by statisticians. Various parametric S-shape models have been proposed for such data. Although they all visually fit well to the PCR data curve, their performances on estimating the initial amount of DNA molecules are not satisfactory. We propose a new approach for the quantification of PCR data to improve the data quality.

Longitudinal Study Of Genome-Wide Dna Methylation From Birth To The First Two Years Of Life

✦Deli Wang, Biostatistics Research Core/CMRC, Children's Memorial Hospital, Northwestern University, 2300 Children's Plaza, Box 205, Chicago, IL 60614 USA, *dwang@childrensmemorial.org*; Xin Liu, Mary Ann and J. Milburn Smith Child Health Program/CMRC, Children's Memorial Hospital ; Ying Zhou, Biostatistics Research Core/CMRC, Children's Memorial Hospital; Xiumei Hong, Mary Ann and J. Milburn Smith Child Health Program/CMRC, Children's Memorial Hospital; Xiaobin Wang, Mary Ann and J. Milburn Smith Child Health Program/CMRC, Children's Memorial Hospital

Key Words: DNA methylation, Genome-wide, DIP test, VMS, Normal mixture, Empirical Bayes

To date, epigenome profiles at birth and in early life are largely unexplored. We performed epigenomic mapping in 105 children(59 boys and 46 girls) enrolled at Boston Medical Center at birth(cord blood) and within the first 2 years of life(venous blood) using Illumina Infinium Human methylation27 BeadChip to address two questions: 1. What is the pattern of genome-wide DNA methylation profiles at birth and in the first 2 years of life? 2. Whether the DNA methylation patterns vary by gender and genomic locations over time. Batch effects were adjusted using an empirical Bayes method.The DIP test,kurtosis and skewness were used to classify methylation distributions. A normal mixture model was applied to classify varied methylated sites(VMS). We identified six major types of distributions of the 27K probes and found significant genome-wide methylation changes within the first two years of life throughout the genome.However,the changes were not significantly different between boys and girls except for X chromosome. We also found that methylation levels and longitudinal changes vary by CpG island and gene structure.We identified top 50 probes which have important biological function.

Conditional Empirical Likelihood Approach To Unbalanced Longitudinal Data Analysis

✦Peisong Han, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, *peisong@umich.edu*; Peter Song, University of Michigan; Lu Wang, University of Michigan

Key Words: GEE, marginal models, robustness

We propose a conditional empirical likelihood approach to analyzing unbalanced longitudinal data, in which stratification is invoked to deal with unbalanced visit patterns. Unlike the currently popular marginal modeling approaches (e.g. GEE), the proposed approach does not require any explicit specification of variance-covariance matrix, but

only correct specification of the marginal mean model. As a result, our approach is robust against model misspecifications in the aspects of marginal variances and/or within-subject correlations. We show that our estimator is consistent and asymptotically normally distributed under certain regularity conditions. In addition, utilizing the objective function in the proposed approach, we establish a likelihood-ratio type of test, which resembles, in both forms and asymptotic properties, to the classical likelihood ratio test. We conduct simulation studies to compare the proposed method with some popular marginal modeling approaches. We also illustrate this method through real data analysis.

Using Markov Chain Composite Likelihood To Analyze Long Sequence Data

✦Jianping Sun, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M2-B500, P.O. Box 19024, Seattle, WA 98109, *jsun@fhcrc.org*; Bruce George Lindsay, Penn State University

Key Words: Markov Chain, Composite Likelihood, Long Sequence Data, Mutation, Recombination

The primary goal of this talk is the analysis of long sequence data generated in biology, such as SNP data. Suppose we have observed n current descendant sequences of length L, one interesting question is that how to estimate the unknown ancestral distribution from the observed descendants, considering realistic biology complexities such as mutation and recombination. We have developed a statistical model with both mutation and recombination to estimate the ancestral distribution. However, though we can write out the full likelihood for ancestral distribution explicitly, there is an enormous computation challenge when applying it on data due to an enormous number of recombination possibilities, which grows exponentially in sequence length. Therefore, we apply composite likelihood as an approximation to solve the problem. In this talk, we first introduce our developed statistical model and composite likelihood method. Then, a Markov chain composite likelihood (MCCL) method is proposed and applied to our statistical model. Finally, some simulation results are shown to investigate the performance of the MCCL.

Robust Linear Regression Methods In Association Studies

✦Vanda Milheiro LourenÁo, CEMAT, IST, Tecnhical Univ. Lisbon, Av. Rovisco Pais, 1, Lisbon, 1049-001 Portugal, *vmilheiro@gmail.com*; Ana Maria Pires, CEMAT, IST, Tecnhical Univ. Lisbon; Matias Kirst, School of Forest Resources and Conservation, University of Florida

Key Words: Non-normality, Outlier, SNP, Power, M-regression, Least squares

Data normality is a mathematical convenience. In practice, experiments usually yield data with nonconforming observations. In the presence of this type of data, classical least squares statistical methods perform poorly, giving biased estimates, raising the number of spurious associations and often failing to detect true ones. Robust statistical methods are designed to accommodate certain types of data deficiencies, allowing for reliable results under various conditions. We analyze the case of statistical tests to detect associations between genomic individual variations (SNPs) and quantitative traits when deviations from the normality assumption are observed. We consider the classical ANOVA tests for the parameters of the appropriate linear model and a robust version of those tests based on M-regression. We show through a simulation study and a real data example, that the robust methodology can be more powerful and thus more adequate for association studies than the classical approach.

302 Synthetic Data, matching and genetic algorithms

Section on Health Policy Statistics

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Prediction Intervals For The Generalized Linear Mixed Model'

✦chenghsueh yang, UCR, 1456 everton pl, riverside, CA 92507, *cyang007@ucr.edu*

Key Words: best linear predictor, best linear unbiased predictor, generalized linear mixed model, pseudo likelihood, prediction interval, quadrature

Some illustrative generalized linear models are used to evaluate the performance (in terms of the coverage probability and the expected width) of alternative prediction intervals for the random effects. A prediction interval based on the eBLUP (empirical Best Linear Unbiased Predictor) that results from the Pseudo-Likelihood (PL) method is a commonly used interval. However, convergence of PL is sometimes not achieved. Alternatively, use of quadrature rules to fit the model is sometimes more viable and does not require as many assumptions as the PL method. The parameter estimates based on the quadrature method can be used to create a prediction interval based on Bayes or best linear predictors. There is an appreciable computational advantage associated with using best linear predictors.

Response Rates When The Cps Follows The Acs

✦Bonnie Coggins, United States Census Bureau, 4600 Silver Hill Road, Suitland, MD 20233, *bonnie.coggins@census.gov*

Key Words: CPS, ACS, date shifting, response rate

It is possible for a household to be selected for both the Current Population Survey (CPS) and the American Community Survey (ACS). If surveys are scheduled within 4 months of each other, the ACS survey date is shifted to allow at least 4 months between the ACS and the CPS interviews. Because these are time-intensive surveys, respondents receiving both surveys may exhibit increased nonresponse for the second survey administered, with higher nonresponse the closer in time the surveys are administered. The ACS is mandatory, but the CPS is not, so we might expect for the CPS to experience more pronounced changes in response rates. This analysis will include data from 2005 through 2009. For CPS respondents receiving the ACS first, various noninterview types will be analyzed and presented graphically and numerically. In addition, a regression of CPS refusal rate vs. lag time between surveys will be calculated. These results will be used to assess the current date shifting strategy for coordinating the CPS and the ACS.

On The Stability Of Sequential Monte Carlo Methods In High Dimensions

✦Alexandros Beskos, University College London, Gower Street, London, WC1E 7HB UK, *alex@stats.ucl.ac.uk*

Key Words: Sequential Monte Carlo, High Dimensions, Markoc chain Monte Carlo

We investigate the stability properties of a class of Sequential Monte Carlo (SMC) methods in high dimensions. It is known in the literature that standard SMC algorithms suffer from curse of dimensionality and can yield computational costs that scale exponentially with the dimension, say d, of the underlying state space. We investigate analytically the properties of an advanced SMC method that develops a sequence a artificial densities between the target and a user-specified proposal, and obtain analytical results for its behavior in high dimensions. In particular, we establish that the cost of the SMC method can be reduced to become quadratic (d^2) with respect to dimension. In essence, we show that SMC methods can be relevant even for high-dimensional scenaria with important implications for the applicability of such methods in practical problems.

A Genetic Algorithm Approach To Optimize Planning Of Food Fortification Programs

✦Dave Osthus, Iowa State University, , *dosthus@iastate.edu*; Alicia Carriquiry, Iowa State University; Todd Campbell, Iowa State University

Key Words: Genetic algorithm, optimization, nutrition, food fortification, measurement error

Methods for reliably estimating the distributions of usual (long-run average) daily nutrient intakes have been proposed (National Research Council, 1986; Nusser et. al. 1996). These estimates are then utilized for evaluating the adequacy of nutrient intake in sub-populations and for the subsequent development of programs to combat those inadequacies. One potential population-level intervention to reduce the prevalence of inadequacy is food fortification - where specific amounts of a nutrient is added to specific food vehicles. The goal of food fortification is to reduce the proportion of the population with inadequate nutrient consumption, at a reasonable cost. But how are food vehicles and nutrient amounts selected? We propose a method to optimize the process to plan food fortification. The approach we propose relies on the methodology proposed by Nusser et al., and uses a genetic algorithm to minimize the analytically untractable optimization function. The goal of the methodology is not to deliver a single, universal "best" food fortification plan, but rather a "best" plan under a variety of constraints. We illustrate the method by planning vitamin A intakes of Ugandan children.

Testing For Absence Of Efficient Combination Gains

✦Pablo Pincheira, Central Bank of Chile, Agustinas 1180, Third Floor, Santiago, International 8320000 Chile, *ppinchei@bcentral.cl*

Key Words: Encompassing, Mean Squared Prediction Errors, Forecast Evaluation

Traditional encompassing tests evaluate if a combination between two forecasts is able to display lower MSPE than the best of the individual forecasts. Nevertheless, it is possible that reductions in MSPE are achieved at the cost of efficiency, which means that these reductions are not necessarily associated with greater forecasting power. In this article we develop a test to evaluate the null hypothesis of absence of efficient combination gains. Rejection of this null implies that it is possible to find a combination between two competing forecasts without paying an efficiency cost. We illustrate the use of this test in an empirical evaluation aimed at examining if efficient combination gains are possible when combining different inflation forecasts for the US and Chile.

A Review Of Disclosure Limitation Methods Employed By Web-Based Data Query Systems

Gregory J. Matthews, University of Connecticut; ✦Robert H. Aseltine, Jr., University of Connecticut Health Center, 263 Farmington Avenue, MC 3910, Farmington, CT 06030, *aseltine@uchc.edu*; Ofer Harel, University of Connecticut

Key Words: Health Policy, Public Health, Statistical Disclosure Control

Public health research often requires access to individuals' health information.††Therefore, data collecting organizations, such as state public health and human service agencies, seek to make these data available while preserving individual privacy for legal and ethical reasons. †One common way in which these data are released is through the use of Web-based Data Query Systems (WDQS). An analysis of the disclosure control techniques used in WDQS recognized by the†National Association for Public Health Statistics and Information Systems†(www.naphsis.org) shows that many state public health query systems fail to employ any type of statistical disclosure control (SDC) and many that do, do so inadequately. In this presentation we seek to stimulate awareness of the potential for privacy breaches via WDQS and offer concrete guidelines for dealing with this problem preemptively, rather than after a major disclosure takes place. †This research is particularly timely in the context of federal healthcare reform, as the emergence of Health Information Exchanges raise the potential for access to vast amounts of de-identified patient data that must be protected to ensure patient privacy.

303 Variable Selection and Dimension Reduction in Nonparametric Models ●

Section on Nonparametric Statistics

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

A Novel Nonparametric Variable Selection For Ultra High Dimensional Problems

✦subhadeep mukhopadhyay, Texas A&M, 415 nagle, apt 8, College Station, TX 77843, *mesuvodeep@gmail.com*

Key Words: high dimension, classification, comparison density, mid-distribution score function

In contrary to the standard practice of variable selection using penalized likelihood method, in this paper we introduce a fast Nonparametric classifier-independent feature selection algorithm for high

dimensional classification problem using the concept of comparison density, mid-distribution score function. The main benefits of our proposed method (i) robust nonparametric model-free, (ii) can detect nonlinear relationship between response and covariates, (iii) can handle continuous, discrete as well as categorical predictor variables, and (iv) can give a deeper insight on the questions 'why`` and 'when`` as well as 'how`` a variable could be influential through visualization. We find that efficiency and effectiveness of our methods and offers additional interpretability and understanding into the operation of variables. Yet surprisingly all available methods are silent on this point. This extra piece of information can help applied researchers not only to identify influential variables but to identify the distinctive characteristics of the variable which makes it so useful for classification - an insightful data analytic variable selection.

Variable Selection In Nonparametric Statistics

✦Adriano Zanin Zambom, Penn State University, 326 Thomas Building, University Park, PA 16803 United States, *adriano.zambom@gmail.com*; Michael G. Akritas, Pennsylvania State University

Key Words: nonparametric regression, ANOVA, variable selection, reduce dimension, asymptotic normality, hypothesis test

In nonparametric regression where we do not want to make restrictive assumptions about the mean function, reducing the dimension of the explanatory variable leads to easier interpretation of the model and better estimates. In this context, we propose a procedure for testing that the nonparametric regression function depends only on a subset of the available covariates, when the mean regression function is not necessarily additive. This hypothesis test is based on recent developments of the asymptotic theory of ANOVA when the number of factor levels goes to infinity. The asymptotic distribution of the test statistic under the null hypothesis is proven to be normal. Simulation results show that this test has better power than previous methods by Lavergne 2000 and Fan and Li 1996, both under linear and nonlinear alternatives.

Additive Partially Linear Regression

✦Fengrong Wei, University of West Georgia, 30117 USA, *fwei@westga.edu*

Key Words: semiparametric, Lasso, consistency

The problem of simultaneous variable selection and estimation in partially linear additive models with a large number of grouped variables in the linear part and a large number of nonparametric components will be considered. In the problem, the number of grouped variables may be larger than the sample size, but the number of important groups is ``small'' relative to the sample size. The adaptive group Lasso is applied to select the important groups, using spline bases to approximate the nonparametric components and the group Lasso is applied to obtain an initial consistent estimator. Under appropriate conditions, it is shown that, the group Lasso selects the number of groups which is comparable with the underlying important groups and is estimation consistent, the adaptive group Lasso selects the correct important groups with probability converging to one as the sample size increases and is selection consistent. The results of simulation studies show that the adaptive group Lasso procedure works well with samples of moderate size.

Local Slicing And A Combining Algorithm On Single-Index Models

✦Wei Lin, Ohio University, Athens, OH 45701, *linw@ohio.edu*

Key Words: single-index model, dimension reduction, nonparametric statistics, local slicing, slicded-inverse regression

The class of single-index models (SIMs) has become an important tool for nonparametric regression analysis. Numerous number of methods and estimators have been developed for the SIM in the past two decades. Each one has its own advantages and disadvantages. They all work well in some cases but poorly in others. On the other hand, many methods utilize an iterative procedure and thus require a good initial point (say, in the neighborhood of $\te_0$ of size n^{-a}). While theoretically existing, such an initial point is not guaranteed by any existing method in practice. Therefore we need a method that can take advantage of multiple existing methods and construct a combined estimator that is more reliable in practice than any single existing one. In this work we present two local slicing methods, both of which utilize a combining algorithm that takes multiple estimators into account. The final estimator is obtained after applying a filtering process with appropriate criteria. Asymptotic results will be given, and a simulation study will be presented which demonstrates that our methods give overall superior results than each individual estimator alone.

On Partial Sufficient Dimension Reduction

✦Xuerong Wen, Missouri University of Science and Technology, 400 W. 12th St., Dept of Math and Statistics, Rolla, MO 65409 USA, *wenx@mst.edu*; Lixing Zhu, Hong Kong Baptist University; Becky Feng, Hong Kong Baptist University

Key Words: sufficient dimension reduction, partial central subspace, partially linear single-index model

Under the general framework of partial sufficient dimension reduction (Chiaromonte et al. 2002), we generalize the notion of partial central subspace from categorical W to continuous ones. Asymptotic properties and small sample properties are also studied. Our method provides a general solution to partial dimension reduction when it is more desirable to conduct dimension reduction on part of the predictors (X) while incorporating the prior information from W (whether W is categorical or continuous), rather than to treat all components of the predictors (X,W) indiscriminately. One immediate application is to the well-known partially linear single-index or multiple-index model (Wang et al. 2010, Carrol et al. 1997).

Hyperplane Alignment: Its Implementation, Application, And Advantages

✦Andreas Artemiou, Michigan Technological University, , *aartemio@mtu.edu*; Bing Li, The Pennsylvania State University; Lexin Li, Department of Statistics. North Carolina State University

Key Words: sufficient dimension reduction, Support vector machines, elliptically distributed predictors, Inverse regression, robustness

Hyperplane alignment (HA) is a new method for sufficient dimension reduction which can effectively extract linear and nonlinear features in the predictor. In this presentation we focus on the implementation and advantages of this method for linear feature extraction through simulation experiments and real data analysis. Since Hyperplane alignment is

based on support vector machine instead of inverse sample moments, it is more robust than traditional dimension reduction methods both against outliers and against non-elliptical distribution of predictors. Furthermore, we demonstrate that it performs well for categorical predictors. An example using the localization sites of proteins in E. coli cells will be discussed

Streaming Algorithms And Their Applications To Hd-Mfpca Models

✦Vadim Zipunnikov, Johns Hopkins School of Public Health, 615 N. Wolfe, E3136,, Department of Biostatistics, Baltimore, MD 21205, *vzipunni@jhsph.edu*; Brian Caffo, Johns Hopkins Department of Biostatistics; Ciprian Crainiceanu, Johns Hopkins University

Key Words: MFPCA, SVD, streaming, MRI

Multilevel Functional Principal Component Analysis for High Dimensional (HD-MFPCA) data combines powerful data compression techniques and statistical inference to decompose the observed data in population- and visit-specific means and subject-specific within and between level variability. However, HD-MFPCA is computationally restricted to the observational studies with a few thousands of observations. We will show how streaming algorithms can be used to overcome this restriction. The suggested algorithm accumulates the information in a streaming fashion resulting in a linear complexity with respect to the sample size. It allows to extend HD-MFPCA to very large samples.

304 New Methods of Estimation

Section on Statistical Computing, International Indian Statistical Association, Section on Statistical Graphics, Section for Statistical Programmers and Analysts

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Quadratic Discrimination for Multi-Level Multivariate Data with Separable Means

✦Anuradha Roy, The University of Texas at San Antonio, Department of Management Science and Statistics, One UTSA Circle, San Antonio, TX 78249, *Anuradha.Roy@utsa.edu*; Ricardo Leiva, F.C.E., Universidad Nacional de Cuyo

Key Words: covariance structure, maximum likelihood estimates, separable means

Under the assumption of multivariate normality we study a quadratic discriminant function of multiple m-variate observations over u-sites and over v-time points. We assume that the m-variate observations have a "jointly equicorrelated covariance" structure and a separable mean vector. A discriminant function is also developed for unstructured mean vectors. The new classification rules are very efficient in discriminating individuals when the number of observations is very small, and thus unable to estimate the unknown variance-covariance matrix. We demonstrate the classification rules on a real data set. Our result shows that our new classification rules are far better than the traditional classification rules for small to moderate sample sizes. These classification rules have plenty of applications in biomedical, medical, pharmaceutical and many other research areas.

Approximating Moments Of An Estimator Of Mean In Ig Populations

✦Debaraj Sen, Concordia University, 1455, De maisonneuve West, Montreal, QC H3G 1M8 Canada, *sen@mathstat.concordia.ca*; Yogendra P. Chaubey, Concordia University

Key Words: Inverse Gaussian Population, Edgeworth expansion, Taylor's series expansion, Least squares, Weighted least squares

This article deals with approximation of the moments of an estimator of the mean of an inverse Gaussian population. Here, we follow the method used in Chaubey and Srivastava (1996) to develop approximation to the first four moments and compare it with those obtained using Taylor series approximation method. The form of the approximation is used in developing empirical formulae as polynomials in the ratio of the square of the coefficient of variation and the sample size.

Moment Estimation Based On Quantiles

✦Haobo Ren, Regeneron Pharmaceuticals, Inc., 400 Somerset Corporate Boulevard, Suite 601, Bridgewater, NJ 08807, *haobo.ren@regeneron.com*; Weining Shen, North Carolina State University; Richard Wu, Regeneron Pharmaceuticals, Inc.; Yuhwen Soo, Regeneron Pharmaceuticals, Inc.

Key Words: Quantile, Mean, Standard Variance, Generalized Lambda Distribution

Quantiles are the critical characteristics of a distribution and play a fundamental role in statistics. However, under some circumstances, we wish to estimate some moment parameters of a sample, such as mean and variance, based on the given empirical quantiles. For example, we were trying to calculate the sample size of a planned clinical trial to compare two-sample means based on a reference article which only provides key quartiles. This experience encouraged us to formulate it as a research topic. Two contributions toward the solution are reported in this talk, one is the derivation of the range of both mean and standard deviation by an optimization approach, the other is to estimate mean and standard deviation by fitting the generalized lambda distributions of quantiles. We focus on the computational perspective and some examples are illustrated.

Approximating Quantiles In Massive Data Sets

✦Reza Hosseini, Simon Fraser University, 8888 University drive, Burnaby, BC V5A 1S6 Canada, *reza1317@gmail.com*

Key Words: quantile, massice data set, approximation, sorting

Very large data sets are often encountered in climatology, either from a multiplicity of observations over time and space or outputs from deterministic models (sometimes in petabytes = 1 million gigabytes). Loading a large data vector and sorting it, is impossible sometimes due to memory limitations or computing power. We develop an algorithm to approximate quantiles of very large datasets which works by partitioning the data or use existing partitions (possibly of non-equal size). We show the deterministic precision of this algorithm and how it can be adjusted to get customized precisions.

Application Of Evolutionary Algorithms In Estimation Of Empirical Likelihoods

✦Ashley Askew, University of Georgia, 101 Cedar Street, Office 263, Athens, GA 30602, *aaskew@uga.edu*

Key Words: Evolutionary algorithm, Empirical Likelihood

An evolutionary algorithm (EA) is a flexible global optimization routine, as well as an alternative to the conventional numerical methods based on derivatives. Empirical likelihoods (ELs) are a non-parametric formulation that utilizes the data to construct a likelihood. The estimation of ELs can pose challenges, especially when constraints must be incorporated in the likelihood. In this talk, we compare the performance of a penalized EA against a numerical procedure based on Newton-Raphson in estimating various examples of ELs.

Parameter Estimation Of Frechet Distribution On Type Ii Censored Data Using Em Algorithm

✦Benhuai Xie, Takeda Global R & D, INC, Deerfield, IL 60015, *bxie@tgrd.com*

Key Words: Tpye-II censoring, Frechet Model, EM algorithm

We consider the estimation of parameters for exponential Frechet distribution under Tpye-II censoring, where the number of units removed at each failure time has a binomial distribution. EM algorithm is applied to obtain Maximum Likelihood estimates. Numerical results are also provided for Tpye-II censoring in Frechet Model.

A Generic Algorithm For Reducing Bias In Parametric Estimation

✦Ioannis Kosmidis, University College London, Department of Statistical Science, University College, Gower Street, London, International WC1E 6BT United Kingdom, *ioannis@stats.ucl.ac.uk*

Key Words: Adjusted score, asymptotic bias correction, bias reduction, fisher scoring, beta regression

A general iterative algorithm is developed for the computation of reduced-bias parameter estimates in regular statistical models through adjustments to the score function (Kosmidis & Firth, 2010, Electronic Journal of Statistics). The algorithm unifies and provides appealing new interpretation for iterative methods that have been published previously for some specific model classes. The new algorithm can usefully be viewed as a series of iterative bias corrections, thus facilitating the adjusted score approach to bias reduction in any model for which the first-order bias of the maximum likelihood estimator has already been derived. The method is tested by application to a logit-linear multiple regression model with beta-distributed responses; the results confirm the effectiveness of the new algorithm, and also reveal some important errors in the existing literature on beta regression.

305 Section on Survey Research Methods - Assessing and Adjusting for Nonresponse

Section on Survey Research Methods, Section on Government Statistics

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Estimation With Non-Ignorable Missing Covariates

✦Jiwei Zhao, University of Wisconsin Madison, 53706, *zhaoj@stat.wisc.edu*; Jun Shao, University of Wisconsin

Key Words: Non-ignorable, missing mechanism, missing covariate, pseudo-likelihood, identifiability, asymptotic normality

Estimation based on data with non-ignorable missing covariates is considered in this paper. We first consider the case where the missing mechanism is nonparametric and the relation between the response and the covariates is of a parametric regression form. We propose a consistent estimator to the regression coefficient when the covariates have non-ignorable missing values. We then show that this estimation procedure can be applied to the case even when the response also has non-ignorable missing values. The identifiability and the asymptotic normality are also derived. Simulation studies are conducted to show the finite sample performance of the proposed estimators. A real data set from the National Health and Nutrition Examination Survey is analyzed.

Propensity Score Adjustment For Nonignorable Nonresponse

✦Minsun Kim Riddles, Department of Statistics, Iowa State University, 1121 Snedecor Hall, Ames, IA 50011-1210, *mskim@iastate.edu*; Jae-kwang Kim, Iowa State University

Key Words: missing data analysis, nonparametric prediction, generalized least squares, EM algorithm

The propensity score adjustment method is commonly used to adjust the bias that is due to nonresponse. We consider the propensity score adjustment method under nonignorable nonresponse. The method we propose does not use a full parametric distributional assumption, but it leads to consistent estimation of the parameters with some moment assumptions. We used the generalized least squares method to combine the observed information and compute an optimal estimator. Variance estimation is discussed, and results from limited simulation studies are presented to show the performance of the proposed method.

Assessment Of Non-Response Bias Through Interviewing Effort Analysis In A Dual Frame Rdd Telephone Survey

✦Wei Zeng, NORC at the University of Chicago, IL 60603, *ZENG-WEI@NORC.ORG*; Ben Skalland, NORC at the University of Chicago; James Singleton, Centers for Disease Control and Prevention

Key Words: Cell Phone survey, non-response bias, weighting

In RDD telephone surveys, direct approaches to assess non-response bias are typically not possible because little information is known about non-respondents. Information linked to landline telephone exchanges do not exist for cell phone samples. An indirect way to evaluate non-response bias in landline and cell phone surveys is a level of interviewing effort analysis, assuming that high-effort respondents may resemble non-respondents. For example, converted refusal cases can be compared to non-refusal cases and respondents can be compared by number of calls made. The National Immunization Survey (NIS)-a nationwide, list-assisted RDD survey, monitors the vaccination rates of children 19 through 35 months. A national cell phone sample for NIS was conducted, targeting households with age-eligible children that only or mainly used their cell phones. This paper uses level of effort analysis to assess the potential non-response bias in the estimates of vaccination coverage in the NIS cell sample, and compares the findings with similar analysis of the NIS landline sample. We will also evaluate whether appropriately weighting the sample can be effective in reducing the potential bias.

Nonresponse Adjustment For A Vector Of Outcomes And The Mar Assumption

✦Joel Wakesberg, Westat, 1600 Research Blvd, Rockville, MD 20850, *joelwakesberg@westat.com*; Karen Christine Masken, IRS Research, Analysis, and Statistics ; J. Michael Brick, Westat Inc.

Key Words: nonresponse, missing data mechanism, vector of outcomes, raking

The IRS National Research Program conducts annual studies of individual taxpayer compliance based on a stratified random sample. As with most studies, not all of the selected taxpayers participate in an audit, resulting in nonresponse that could be due to various reasons including missing returns, unlocatable taxpayers, or taxpayers who never respond to any IRS correspondence. Historically, the IRS has adjusted the respondents weights using assumptions about the missing data mechanism based on the reason for nonresponse. This approach deals with one important dimension and addresses total noncompliance, the key estimate from the survey. However, some single line item entries such as self-employed business income or charitable contributions are of great interest to analysts as well and the adjustment may be less effective for these items. In our paper, we review the rationale for treating all nonresponse as missing at random. We then explore ways to adjust for the nonresponse bias when analysts are interested in a vector of estimates, not just one point estimate. Finally, we demonstrate that raking to multiple variables is very effective as compared to the traditional approach.

Analysis Of Nonresponse In The Statistics Of Income'S 1999 Individual Tax Return Panel

✦Tara R Wells, Statistics of Income / Internal Revenue Service, 77 K Street NE, Washington, DC 20002, *tara.r.wells@irs.gov*

Key Words: Nonresponse bias, Propensity score adjustment, Panel attrition

In 1999, the Statistics of Income (SOI) Division of the Internal Revenue Service began collecting individual tax returns for SOI's 1999 Individual Tax Return Panel. Longitudinal individual tax data is essential to study how taxpayers react to tax law changes and how the tax system affects taxpayers over an extended period of time. However, as in all panels, SOI's 1999 Individual Tax Return Panel is impaired by panel attrition. Previous papers have evaluated the presence and motivation for attrition in prior SOI individual tax return panels, yet there has not been any research that has measured the nonresponse error caused by the panel attrition. In this research, I use an exploratory approach to estimate the nonresponse error in specific tax-related variables collected from SOI's 1999 Individual Tax Return Panel from 2000 to 2003. Then I use a propensity score method of subclassification to investigate if the nonresponse bias in the SOI 1999 Individual Tax Return Panel can be removed. My results show that the nonresponse bias in the tax-related variables can be reduced with the use of propensity score adjustments.

Representativeness (R-Index) And Nonresponse Bias Patterns In Household Surveys

✦John Dixon, Bureau of Labor Statistics, 2 Massachusetts Ave, NE, Room 1950, Washington, DC 20212, *dixon_j@bls.gov*

Key Words: Survey Nonresponse, R-index, propensity scores

Nonresponse rates have been used as a proxy for survey quality since they indicate the relative potential for nonresponse bias. Recently the R-index (Schouten) has generated interest in an alternative approach that better represents the potential for bias by focusing more on coverage than nonresponse. The patterns of nonresponse rates (e.g.; seasonal, time in sample) and the R-index can provide insight into the usefulness of nonresponse rates and representativeness. The current study uses different measures of nonresponse bias, nonresponse rates, and the R-index to see if there are patterns for bias and representativeness which might be different than for response rates alone. Two surveys, the Current Population Survey (CPS), and the Consumer Expenditure Quarterly Survey (CEQ) will be used in this analysis.

Reweighting In The Presence Of Nonresponse In Stratified Designs

✦Ismael Flores Cervantes, Westat, 1600 Research Blvd, Rockville, MD 20850, *ismaelflorescervantes@westat.com*; J. Michael Brick, Westat Inc.

Key Words: Nonresponse, stratification, weighting classes

Reweighting a sample using weighting class adjustments is one approach to deal with nonresponse. This approach uses a response model defined as a set of assumptions about the true but unknown response distribution that corresponds to the weighting class. A reweighted estimator is unbiased if the model coincides with the response distribution. However, in most cases, the response model will differ from the true response distribution. In this paper we examine the effect of reweighting when the model fails in stratified designs. The majority of results on model failure in nonresponse in the literature assume a simple random sampling. We expand this to stratified designs and compare the results with other approaches such as nonresponse adjustments that ignore sampling weights.

306 Diagnostic Testing and Computational Analysis ■

Biopharmaceutical Section

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Standard Errors And Confidence Intervals For Target Dose Estimation Under Model Uncertainty

✦Shanhong Guan, Merck & Co., 351 N. Sumneytown Pike, North Wales, PA 19454, *shanhong_guan@merck.com*; Jose Carlos Pinheiro, Johnson & Johnson PRD

Key Words: Bootstrap, Dose-finding, Miminum effective dose, Multiple comparison procedure

Identifying and estimating target doses, such as the smallest dose achieving a clinically relevant effect, (minimum effective dose, MED), is a key goal in drug development. A unified approach, combining multiple comparison procedures and modeling (MCP-Mod), was proposed by Bretz, Pinheiro and Branson (2005) for designing and analyzing dose finding studies. In this paper, we describe several bootstrap procedures for determining the precision of the MED estimates using the MCP-Mod methodology, including the corresponding standard error and the 90% confidence interval. The results are compared with available asymptotic confidence intervals via an extensive simulation study. Applications of these methods are illustrated using a phase II trial.

Evaluation Of Statistical Methods For Sample And Diagnostics Test Stability

✦Zhonggai Li, Novartis Molecular diagnostics (MDx), 45 Sidney St, Cambridge, MA 02139, *zhonggai@gmail.com*; Holger Hoefling, Novartis Molecular Diagnostics (MDx); Anthony Rossini, Novartis Molecular Diagnostics (MDx)

Key Words: sample stability, diagnostic test comparisons, constant bias, archived samples, personalized medicine

A common statistical quantity that must be assessed in the development of clinical diagnostics is the stability of results over time. The concerns over stability of diagnostic results can be due to clinical sample storage, change in reagent batches, change of assay machines, introduction of new technologies into the protocol for scientific or economic reasons, or similar changes. Our research addresses the selection of statistical approaches to resolve the specific question "how can we evaluate whether the storage of a clinical sample biases the stability of the diagnostic readout?". We focus on the evaluation of constant bias/difference which could be due to sample storage or changes between a new and original test. These two situations often occur during development of personalized medicine diagnostics as a result of the limited availability of precious clinical samples and desires to leverage newer technology for existing diagnostics for scientific or economic reasons. We offer a summary report on the pros and cons of each evaluated statistical test and recommend appropriate statistical tests for assessing bias in the context of stability of new tests or archived samples.

Verification Bias In Diagnostic Devices Submissions

✦Marina V. Kondratovich, U.S.FDA, 10903 New Hampshire Avenue, White Oak 66, Room 5666, Silver Spring, MD 20993, *Marina.Kondratovich@fda.hhs.gov*

Key Words: diagnostic device, study design, verification bias

An issue of verification bias presents in a lot of submissions to the FDA. If it is not addressed properly by an appropriate study design or an appropriate type of statistical analysis, it can create problems (sometimes unfixable) for unbiased evaluation of medical tests. We present different schemes of verification bias along with examples: as 1) a random sample of the subjects with negative results via both tests (Old test and New test) has gold standard results; 2) no subjects with negative results via both tests have gold standard results; we describe two schemes: one for a paired study design (pre-market studies) and another for an unpaired study design (post-market studies); 3) only subjects with positive results by the Old Test have results of the gold standard and were included in the study; 4) only the subjects with the Old test positive results have results of the gold standard in a study with follow-up; 5) example of verification bias of the test for drug response.

Application of the EM Algorithm for Mixtures of Distributions with Added Information

✦Chen Teel, Dupont USA, 117 Presidential Drive, Apt B, Greenville, DE 19807, *gcheer3@gmail.com*; Taeyoung Park, Yonsei University; Allan Sampson, University of Pittsburgh

Key Words: EM algorithm, mixtures of distributions, conditional Bernoulli distribution, exponential family

The EM algorithm has been widely used to estimate parameters in mixtures of distributions where parameters in each component distribution are unknown and mixing proportions may be known or unknown. Here we consider such mixtures of distributions, but with added information as to mixture components. In particular, the mixtures of two exponential family distributions are considered when the number of observations within each mixture component is known. This situation frequently occurs in an adaptive design setting where block randomization are often used in clinical trials. By fully capitalizing on the additional information as to mixture component size, we develop a new computational method to implement the EM algorithm for mixtures of distribution. Our algorithm shows robustness to the choice of starting values and exhibits a fast and stable convergence property.

Graphic Display For Summarizing Individual Responses In Crossover Designed Human Abuse Potential Studies

✦Ling Chen, FDA/CDER, 10903 New Hampshire Ave., WO21 RM 4644, Silver Spring, MD 20903, *ling.chen@fda.hhs.gov*

Key Words: Abuse potential, Crossover study, Extreme response, Graphic method, Peak duration

The human abuse potential study plays a critical role in understanding whether a drug produces positive subjective responses indicative of abuse potential. This type of study has crossover designed investigation with multiple treatments and multiple abuse potential measures. Sponsors often provide mean time course profiles for each abuse potential

measure by treatment, but this does not provide information about time to peak or peak duration for individual subjects. This presentation will propose a graphic method to display individual responses in a crossover study. This graphic method will provide an easy tool for the Controlled Substance Staff (CSS) at FDA and Sponsors to visually evaluate whether individual responses to each treatment are different from each other, and also provide a tool to investigate the time of peak response and the duration of the peak response as well as extreme responses on a subject base.

Graphical Analysis Of Safety Data In Nda Submissions

✦Linyun Zhou, Takeda Global Research & Development, One Takeda Parkway, Deerfield, IL 60015, *linyun.zhou@tgrd.com*

Key Words: graphical analysis, relative risk, patient narrative, DILI plots

Patient safety has always been a primary focus in the development of new pharmaceutical products. The predominant method for statistical evaluation and interpretation of safety data collected in a clinical trial is the tabular display of descriptive statistics. There is a great opportunity to enhance evaluation of drug safety through the use of graphical displays, which can convey multiple pieces of information concisely and more effectively than can tables. From some approved NDA submission, this paper provides graphical analysis in some key safety data. The graphical displaying safety data in this paper primarily focus on patient narratives, relative risk and adverse event, DILI plots for liver function, and line plots for multiple laboratory safety parameters and/or vital signs.

Innovative Application Of Statistical Modeling Methods In The Field Of Imatinib Treatment Effects On Chronic Myeloid Leukaemia Patients

✦Min Tang, Harvard University, 02120, *min@jimmy.harvard.edu*; Mithat Gonen, Memorial Sloan-Kettering Cancer Center; Franziska Michor, Harvard University

Key Words: model fit, bootstrapping, Chronic Myeloid Leukaemia

Treatment of Chronic Myeloid Leukaemia (CML) with the tyrosine kinase inhibitor imatinib represents a successful application of molecularly targeted cancer therapy. However, the effect of imatinib on leukaemic stem cells is current unknown. Which model fits better to the clinical data, an exponential curve or biphasic exponential curve with an unknown turning point, will result in completely different biological interpretation and understanding of the imatinib treatment effects on leukaemic stem cells. We propose a statistical test to implicitly test the fit of the two models on the longterm imatinib clinical data and address that important biological question using bootstrapping simulations. We show that successful long-term imatinib treatment results in a biphasic exponential decline of BCR-ABL1 transcript levels starting from the 6th month's treatment. This innovative statistical modeling approach suggests that long-term imatinib treatment has the potential of reducing the number or proliferation ability of leukaemic stem cells.

307 Multiplicity and Multiple Comparisons (II) ■

Biopharmaceutical Section

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Selecting The Superior Dose After An Interim Analysis Taking Into Consideration Type-I Error And Power Configuration

✦Haifeng Zhang, University of Toledo, , *hfzhang3@hotmail.com*

Key Words: Conditional Power, Type I error, interim analysis

This presentation focuses on multiple dose comparisons with a control for the purpose of identifying at least one dose which is superior to a comparator. An interim analysis can be used to drop dose(s) that are found to be either inferior or having little chance to show a statistically significant superiority to the comparator. We study the type-one error and power of various design configurations and propose a strategy for such an analysis.

Optimal Decision Trees In Gatekeeping Procedures For Confirmatory Clinical Trials

✦JonDavid Sparks, Eli Lilly & Company, Indianapolis, IN 46285, *sparks_jondavid@lilly.com*; Alex Dmitrienko, Eli Lilly and Company; Haiyuan Zhu, Forest Research Institute

Key Words: multiple testing, gatekeeping procedures, grouping objectives

Confirmatory clinical trials often consist of many objectives where the need to adjust for multiplicity is required. The objectives are often grouped into families where serial or parallel gatekeeping procedures may be utilized. The decision on how to group and order families (e.g., primary, secondary) may be based on clinical relevance, assumed statistical power, or for various other reasons. However, concerns can arise as the number of hypotheses in a family or the number of families increases. This may lead to a substantial decrease in power for some promising objectives or objectives may simply not be tested due to the performance in previous families. In this research, we explore approaches that group objectives into different families and test them using general gatekeeping procedures. A variety of multiple testing procedures in combination with different family groupings will be considered. The power performances of the strategies are evaluated under various nominal power scenarios. An optimal decision tree is presented which can aid researchers in developing multiple testing methods under assumed nominal power levels.

An Extended Hochberg'S Step-Up Procedure On Testing Two Individual Null Hypotheses

✦Yi Ma, Quintiles, 6700 W 115th St, Overland Park, KS 66211, *yi.ma@quintiles.com*; Huajiang Li, Allergan

Key Words: multiple testing procedure, Hochberg's step-up procedure, overall type I error rate

In this presentation, we will first briefly review several multiple testing procedures (MTP) widely used in clinical trials based on univariate p-value adjustment to strongly control overall type I error rate. These

MTPs include Bonferroni, fixed sequence, fallback, Holm's step down and Hochberg's step up procedure. These procedures on testing two null hypotheses will be illustrated intuitively in diagrams. We will then propose a new method extending Hochberg's step-up procedure in order to improve the testing power and simultaneously strongly control the overall type I error rate. Simulation results will be presented comparing our method and the Hochberg's step-up procedure in situations that individual p-values are independent, positively and negatively correlated.

A New Partition Testing Strategy For Multiple Endpoints

✦Bushi Wang, University of California, Riverside, 3384 Idaho St, Riverside, CA 92507, *bushi.wang@email.ucr.edu*; Xinping Cui, University of California, Riverside

Key Words: Multiple endpoints, Likelihood ratio test, Partition testing, Consonance, Union-intersection test

To evaluate efficacy in multiple endpoints in confirmatory clinical trials is a challenging problem. The difficulty comes from the different importance of each endpoint and their underlying correlation. Current approaches to this problem are based on closed testing or partition testing, which test the efficacy in certain dose-endpoint combinations and collate the results. Despite their different formulations, all current approaches test their dose-endpoint combinations as intersection hypotheses and apply various union-intersection tests. Likelihood ratio test is seldomly used due to the extensive computation and lacks of consistent inferences. In this talk, I will first generalize the decision path principle proposed by Liu and Hsu (2009) to the cases with alternative primary endpoints and co-primary endpoints. Then I will propose a new partition testing approach which is based on consonance adjusted likelihood ratio test. The new procedure provides consistent inferences and yet it is still conservative and does not rely on the estimation of endpoint correlation or independence assumptions which might be challenged by regulatory agencies.

A Natural Approach For Addressing Multiplicity Involving Neuroscience Functional Data

✦Junshui Ma, Merck Research Laboratories, Merck & Co. Inc., 126 E Lincoln Ave, RY33-300, Rahway, NJ 07065, *junshui_ma@merck.com*; Svetnik Vladimir, Merck Research Laboratories, Merck & Co. Inc.

Key Words: Clinical Data Analysis, Neuroscience Data Analysis, Functional Data Analysis, Multiple Comparison, Family Wise Error Rate, High Dimensional Integration

Many neuroscience modalities, e.g. fMRI, and EEG, output functional data in the time, spatial, or frequency domain. Multiplicity issue arises when inferences are simultaneously conduced at discretized domain points. The nature of functional data suggests that the domain should be treated as continuous, and the correlation among the inferences should be considered. The approaches that address multiplicity by adjusting p-values are not applicable. A natural solution based on functional data analysis is proposed. The essence of the proposed approach is to estimate the joint distribution of the simultaneous inferences, and directly calculate the FWER from the joint distribution. When the functional data, along with other factors (e.g. treatment), are modeled with a set of continuous basis functions, and the inferences are represented as linear transformation of the basis functions and model coefficients, the covariance matrix of the inferences can be estimated from the fitted model. Thus, the joint distribution of the inferences is obtained under Gaussian assumption. Calculating FWER from the joint distribution requires high dimensional integration, which recently becomes feasible.

Multiple Endpoints With Latent Variable Approach

✦Juanmei Liu , Medtronic Inc, 94086 USA, *juanmeiliu@yahoo.com*; Minglei Liu, Medtronic Inc

Key Words: multiplicity, multiple endpooints, latent variable, correlation structure, polychoric correlation

Traditional approaches dealing with multiplicity from multiple endpoints are in the form of p-value adjustment.These approaches will not take in account the possible latent structure of the multiple endpoints and may have lower power. To gain power, we developed latent variable approach in theory by considering correlation/covariance structure of the multiple outcomes. With reduced dimension, no adjustment of p values or only less adjustment are required. We developed conditions under which this approach is applied and we generalized this method to incorporate all types of endpoints. A simulation study is presented and an example from real clinical trial data is analyzed and interpreted.

A Consistency-Adjusted Strategy For Testing Alternative Endpoints In A Clinical Trial

✦Mohamed A. Alosh, FDA, 10903 New Hampshire Ave., Bldg.# 21, Room 3626, Silver Spring, MD 20993 USA, *Mohamed.Alosh@fda.hhs.gov*

Key Words: Alternative endpoints, subgroups, consistency, dependency, study power

A clinical trial might involve more than one clinically important endpoint (subgroup) each of which can characterize the treatment effect of the experimental drug under investigation. For prespecifeid alternative endpoints (subgroups) there are several approaches which can be used for testing for efficacy for the alternative endpoints or the subgroup and total study population. Traditional multiplicity approaches use constant significance levels for these alternative endpoints. However, some recent multiplicity strategies allow the alpha-level allocated to testing subsequent endpoint to be dependent on the results of previous endpoint. In this presentation we discuss the need for establishing a minimum level of efficacy for the previous endpoint before proceeding to test for the subsequent alternative endpoint (subgroup) so that potential problems in interpreting study findings can be avoided. We consider implementing such requirements, called consistency criteria, along with adaptation of the significance level for subsequent endpoints at the study design stage and investigate its impact on study power. In addition, we consider its application to actual clinical trial data.

308 Robust Methods and Methods for Heavy Tails

Business and Economic Statistics Section

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Gel Estimation For Semi-Strong Non-Linear Garch With Robust Empirical Likelihood Inference

✦Jonathan Hill, University of North Carolina, Dept. of Economics, Chapel Hill, NC 27599 USA, *jbhill@email.unc.edu*; Artem Prokhorov, Concordia University

Key Words: Empirical Likelihood, Nonlinear GARCH, Tail Trimming, Empirical Likelihood Ratio Test

We construct Generalized Empirical Likelihood estimators for random volatility models of heavy-tailed data with particular attention to nonlinear GARCH. The estimator imbeds trimmed estimating equations allowing for over-identifying conditions, consistency, asymptotic normality and efficiency for very heavy-tailed data due to feedback or idiosyncratic noise. As opposed to existing heavy tailed robust QML and LAD estimators for random volatility (Ling 2005, Peng and Yao 2003, Linton et al 2010) we allow for model asymmetries and over identifying restrictions. We use the theory of GEL with tail-trimming to construct several robust tests that uses as plug-in any consistent estimator of the parameter and likewise characterize a class of efficient moment estimators.

Partially Linear Modeling For Conditional Quantile

✦Chaojiang Wu, University of Cincinnati, Room 534, 2925 Campus Green Drive, Cincinnati, OH 45220, *wuca@mail.uc.edu*; Yan Yu, University of Cincinnati

Key Words: Additive Models, Dimension Reduction, Penalized Splines, Single-Index Models, Smoothing Parameter, Semiparametric Model

We consider the estimation problem of conditional quantile when high demensional covariates are involved. To overcome the "curse of dimensionality" yet retain model flexibility, we propose to two partially linear models for conditional quantile: partially linear single-index models (QPLSIM) and partially linear additive models (QPLAM). The unknown functions are estimated by penalized splines. An iteratively reweighted least square algorithm is developed. To facilitate model comparisons, we develop effective model degrees of freedom as the measure of model complexity for penalized spline conditional quantile. Two smoothing parameter selection criteria, Generalized Approximate Cross-validation (GACV) and Schwartz-type Information Criterion (SIC) are studied. Some asymptotic properties are established. Finite sample properties are studied by simulation studies. A real data application demonstrates the success of proposed approach. Both simulations and real applications show encouraging results of our estimators.

Quantile Autocorrelation Function And Quantile Partial Autocorrelation Function

✦Yang Li, The University of Hong Kong, Department of Statistics and Actuarial Science, Room 518, Meng Wah Complex Building, The University of Hong Kong, Hong Kong, International China, *snliyang@hku.hk*

Key Words: quantile regression, autocorrelation, partial autocorrelation, model specification, diagnostic checking

Since the pioneer research of quantile regression existed in the later decades of last century, lots of its applications followed. With noted that the quantile research on time series mainly focused on model estimation, we reconstruct some traditional statistics to verify their asymptotic properties to check the significance of estimated parameters and the randomness of errors. Encouraged by the will of working out a new system of benchmarks of time series model, we have worked on how the movements of t-th observation were influenced by its forehead observation at lag k. We figure out their by redefining the autocorrelation and partial autocorrelation function of a stationery sequence in quantile ways. We express the quantile autocorrelation function (QACF) and partial autocorrelation function (QPACF), then observe their curtail properties in order to complete the quantile identification process. Based on the new standard model specification, we deduce several new statistics under the quantile parameter estimation method and figure out their asymptotic properties to fill the diagnostic checking part of the quantile regression.

Inference In Predictive Quantile Regressions

✦Alex Maynard, University of Guelph, Department of Economics, , *maynarda@uoguelph.ca*; Katsumi Shimotsu, Department of Economics, Hitotsubashi University; Yini Wang, Department of Economics, Queen's University

Key Words: local-to-unity, quantile regression, predictive testing

This paper studies inference in predictive quantile regressions when the predictive regressor has a near-unit root. We derive nonstandard distributions for the quantile regression estimator and t-statistic in terms of functionals of diffusion processes. The critical values are found to depend on both the quantile of interest and the local-to-unity parameter, which is not consistently estimable. Based on these critical values, we propose a valid Bonferroni bounds test for quantile predictability with persistent regressors. We employ this new methodology to test the ability of many commonly employed and highly persistent regressors, such as the dividend yield, earnings price ratio, book to market ratio, term spread and T-bill rate, to predict the median, shoulders, and tails of the stock return distribution.

Least-Squares Estimation And Order Selection For Heavy-Tailed Arma Time Series With Garch Innovations

✦Huanhuan Wang, Northwestern University, 2006 Sheridan Rd, Evanston, IL 60208, *whhelia@gmail.com*; Beth Andrews, Northwestern University

Key Words: Autoregressive-moving average, GARCH, least-squares estimation, order selection

Since least-squares is a standard preliminary estimation technique, we consider properties of least-squares estimators for autoregressive-moving average (ARMA) time series model parameters when the ARMA process is heavy-tailed with GARCH innovations and tail index in the interval (2,4). These ARMA-GARCH series have infinite fourth but finite second moments, properties exhibited by many observed time series, particularly in finance. In this case, the least-squares estimators of the ARMA model coefficients are consistent and converge in distribution to a function of non-Gaussian stable random variables, with rate of convergence slower than n^{1/2}. Using the asymptotic distribution for the least-squares estimators, we identify information criterion statistics

which can be used for consistent estimation of ARMA model order. We examine finite sample behavior of the ARMA estimators and order selection statistics via simulation, and the techniques are used to fit an ARMA-GARCH model to heavy-tailed financial time series data.

Assessing Extremal Dependence In Equity Markets

✦Jose Faias, Universidade Catolica Portuguesa, Palma de Cima, Lisbon, International 1649-023 Portugal, *jfaias@fcee.ucp.pt*; Miguel de Carvalho, Ecole Polytechnique FÈdÈrale de Lausanne, Swiss Federal Institute of Technology; AntÛnio Rua, Banco de Portugal

Key Words: Asymptotic independence, Multivariate extreme theory, Tail dependence, Empirical likelihood, Risk modelling, Sectoral tail comovements

In recent years there has been an increasing interest in modelling dependence in heavy tail phenomena such as the latest turbulence episodes in financial markets. The evidence of asymptotic independence in the financial data has led to the need of rethinking risk modelling and inference tools for multivariate extremes. In this paper we propose an inference scheme for assessing extremal dependence of several pairs of variables, in the context of asymptotic independence. Our approach is based on the fact that the problem of interest can be rewritten as an empirical likelihood problem for comparing the means of different populations, where such means represent the Hill's estimate of the coefficient of tail dependence. A triangular array representation allow us to obtain a nonparametric Wilks' theorem, and we apply the method to assess extremal dependence in equity markets.

Systematic Risk Under Extremely Adverse Market Conditions

✦Chen Zhou, Erasmus University Rotterdam, P.O. Box 1738, Rotterdam, International 3000DR The Netherlands, *zhou@ese.eur.nl*; Maarten van Oordt, De Nederlandsche Bank

Key Words: systematic risk, tail dependence, extreme value theory, Value-at-Risk

Extreme losses are the major concern in risk management. The dependence between financial assets and the market portfolio changes under extremely adverse market conditions. We develop a measure of systematic tail risk, the tail regression beta, defined by an asset's sensitivity to large negative market shocks, and establish the estimation methodology. Building on extreme value theory, the estimator of the tail regression beta consists of the asymptotic dependence measure and the marginal risk measures. Theoretically, it has a similar structure as the estimator of the regular beta from regression analysis. Simulations show that our estimation methodology yields an estimator that has a lower mean squared error than performing regressions in the tail. Empirical results based on analyzing 46 industrial portfolios demonstrate that the regular systematic risk measure is in general different from the systematic tail risk in severe market downturns. Furthermore, the tail regression beta is a useful instrument in both portfolio risk management and systemic risk management. We demonstrate its applications in analyzing Value-at-Risk (VaR) and Conditional Value-at-Risk (CoVaR).

309 Statistical Inference for Stochastic Processes ●

IMS

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

On The Estimation Of Locally Stationary Processes

✦Wilfredo Palma, Pontificia Universidad Catolica de Chile, , *wilfredo@mat.puc.cl*

Key Words: Locally stationary, Estimation, Consistency, Normality, Simulations, Applications

This talk addresses the estimation of locally stationary processes. A time-varying parametric formulation of these models is discussed and maximum likelihood techniques are proposed for estimating the parameters involved. Large sample properties of some of these estimates such as consistency, normality and efficiency are established. Furthermore, the finite sample behavior of the estimators is investigated through Monte Carlo experiments. As a result from these simulations, we show that the estimates behave well even for relatively small sample sizes. Several real-life data applications are also presented.

On Estimating Threshold Crossing Times

✦Tony Sit, Columbia University, Room 1005, MC4690, 1255 Amesterdam Avenue, New York, NY 10027, *tony@stat.columbia.edu*; Victor de la Pena, Columbia University; Mark Brown, City College, CUNY

Key Words: First-hitting time, Threshold-crossing, Probability bounds, Decopuling, Climate change, Renewal theory

Given a range of future projected climate trajectories that represent the future climate condition, an unbiased estimator for the threshold-crossing time will be the mean of the crossing-times of all the simulated paths. This estimator, however, can provide a sensible estimation only when all the simulated paths have infinite/observed first-hitting times. In this paper, we propose remedies to the estimation problem to deal with situations in which there is one or more simulated paths having a right censored boundary crossing time. This extends the results in Section 2.7 of de la Pena (1997) and provides a universal sharp lower bound than what is shown in de la Pena and Yang (2004). Two examples of applications of the bounds derived are provided: one involving the growth of cancer tumours and the other one deals with drought prediction in US Southwest and Mediterranean region based on the data calculated from IPCC Fourth Assessment (AR4) model simulations of the twentieth and twenty-first centuries.

Estimation Methods for Nonlinear Time Series

✦Candace Metoyer, Intel Corporation, Santa Clara, CA 95054, *candace.n.metoyer@intel.com*; Prabir Burman, University of California, Davis

Key Words: nonlinear time series, state space models, penalized likelihood, asymptotic mean square error, poisson, bernoulli

We consider the class of structural models for nonlinear time series where the underlying signal may contain trend and seasonal components. In particular, we investigate signal estimation methods for time series whose observations come from a distribution that is a member of the exponential family of distributions. Common examples of these data include Poisson time series (which arise from count data) and Bernoulli time series (which arise from binary response data). A method based on penalized log-likelihood is used to generate estimates of the signal components. Asymptotic results for the mean square error of the estimators are given and applications to real time series data are provided.

On The Approximate Maximum Likelihood Estimation For Diffusion Processes

✦Jinyuan Chang, Peking University, Building 26, Room 224, Peking University, Beijing, 100871 China, *changjinyuan1986@yahoo.com.cn*

Key Words: Asymptotic normality, Consistency, Discrete observation, Edgeworth expansion, Maximum likelihood estimation

The transitional density of a diffusion process is generally unknown, which prevents the full maximum likelihood estimation (MLE) based on discretely observed sample paths. A\"it-Sahalia (1999, 2002) proposed Edgeworth type series approximations to the transitional densities of diffusion processes, which lead to the approximate maximum likelihood estimation (AMLE) for parameters. The consistency and the rate of convergence of the AMLE are established, which reveal the roles played by the number of terms used in the density approximation and the sampling length between successive observations. We find conditions under which the AMLE have the same asymptotic distribution as that of the full MLE. A first order approximation to the Fisher information matrix is proposed.

Fast Convergence Rates In Estimating Large Volatility Matrices Using High-Frequency Financial Data

✦Minjing Tao, University of Wisconsin-Madison, , *minjing@stat.wisc.edu*; Yazhen Wang, University of Wisconsin-Madison; Xiaohong Chen, Yale University

Key Words: large dimensional diffusion, matrix norm, micro-structure noise, multi-scale realized volatility matrix estimator, sparsity, threshold

Financial practices often need to estimate an integrated volatility matrix of a large number of assets using noisy high-frequency data. Many existing estimators of volatility matrix of small dimensions become inconsistent when the size of the matrix is close to or larger than the sample size. This paper introduces a new type of large volatility matrix estimators based on non-synchronized high-frequency data, allowing for the presence of micro-structure noise. When both the number of assets and the sample size go to in?nity, we show that our new estimator is consistent and achieves fast convergence rate, where the rate is optimal with respect to the sample size.

On Data Adaptive Wavelet Decomposition And Bootstrap Under Long-Range Dependence

✦Jan Beran, University of Konstanz, Department of Mathematics and Statistics,, University of Konstanz, Universitaetsstrasse 10, Konstanz, International 78457 Germany, *jan.beran@uni-konstanz.de*; Yevgen Shumeyko, University of Konstanz

Key Words: long-range dependence, wavelets, bootstrap, discontinuity, data adaptive, time series

Optimal data adaptive wavelet estimation of a trend function in time series with long-range dependence leads to a natural decomposition into a low and a high-resolution part. Explicit formulas for optimal decomposition levels and smoothing parameters are obtained. Moreover, a bootstrap test is developed to detect discontinuities in the underlying trend function. Asymptotic validity and consistency of the test are derived under general conditions.

Subsampling Weakly Dependent Times Series And Application To Extremes

✦Silika Prohl, University of Zurich and IORFE (Princeton University), Sherrerd Hall, Charlton Street, Princeton, NJ 08544, *silika.prohl@uzh.ch*; Paul Doukhan, Laboratory of Mathematics UCP, University Cergy-Pontoise ; Christian P Robert, Universite Paris-Dauphine

Key Words: Extremes, Subsampling, Weak dependence, Mixing

Politis and Romano (1994) established the subsampling estimator for converging statistics when the underlying sequence is strongly mixing. Bertail \textit{et al.} (2004) applied this work to subsampling estimators for distributions of diverging statistics. In particular, they constructed an approximation of the distribution of the sample maximum without any information on the tail of the stationary distribution. However, the assumption on the strong mixing properties of the time series is sometimes too strong as for the class of first-order autoregressive sequences with uniform marginal distribution introduced and studied by Chernick (1981): let $(X_{t})_{t\in \mathbb{Z}}$ be uniform AR(1) process defined recursively as \begin{equation} X^{(r)}_{t}=\frac{1}{r} X^{(r)}_{t-1}+\varepsilon _{t}, \label{FOS} \end{equation} where $r\geq 2$ is an integer, $(\varepsilon _{t})_{t\in \mathbb{Z}}$ are iid and uniformly distributed on the set $\{0,1,\ldots ,r-1\}$ and $X^{(r)}_{0}$ is uniformly distributed on $[0,1]$. The results of Bertail \textit{et al.} (2004) can not be used for this class of processes although the normalized sample maximum has a non-degenerate limiting distribution

310 Applications of Statistical Graphics

Section on Statistical Graphics, Section on Statistical Computing, Section for Statistical Programmers and Analysts

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Msblffm: Multivariate Semiparametric Bayesian Local Factor Functional Models For Diffusion Tensor Tract Statistics

✦Zhaowei Hua, UNC-Chapel Hill, 27517, *zhua@bios.unc.edu*; David Dunson, Duke University ; Hongtu Zhu, University of North Carolina Department of Biostatistics

Key Words: Bayes confidence band, Diffusion tensor imaging, Fiber bundle, local inference, Dirichlet process

Diffusion tensor imaging (DTI) is a modality to visualize and quantify the structure of white matters in human brain. In this article, we propose a multivariate semiparametric Bayesian local factor functional model to analyze fiber tract data. A local partition process is used to address the variability of multiple diffusion properties along major white fiber bundles and its association with a set of covariates of interests, such as gestational age. Two types of statistical inferences are provided: (1) global hypothesis testing to test the overall significance of a hypothesis of interest (2) local hypothesis testing to identify the region of significance. Posterior computation proceeds via an efficient MCMC algorithm using the exact block Gibbs sampler. A simulation study is performed to evaluate the performance of MSBLFFM. Our method is applied to analyze a fiber track data set of two fiber tracts from a clinical study of neurodevelopment: the splenium of the corpus callosum tract and the right internal capsule tract. The growth of white matter fiber diffusivities along these two tracts are addressed.

Multiscale Adaptive Composite Quantile Regression Models For Neuroimaging Data

✦Linglong Kong, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514 USA, *lkong@bios.unc.edu*; Hongtu Zhu, University of North Carolina Department of Biostatistics

Key Words: Kernel, Multiscal adaptive regression, Neuroimaging data, Propagation separation, composite quantile regression, Robustness

Neuroimaging studies aim to analyze imaging data with complex spatial patterns in a large number of locations (called voxels) on a two-dimensional (2D) surface or in a 3D volume. We propose a multiscale adaptive composite quantile regression model (MACQRM) that has four attractive features: being robustness, being spatial, being hierarchical, and being adaptive. MACQRM utilizes imaging observations from the neighboring voxels of the current voxel and borrows strength from the nearby quantile regressions of the current regression to adaptively calculate parameter estimates and test statistics. Theoretically, we establish consistency and asymptotic normality of the adaptive estimates and the asymptotic distribution of the adaptive test statistics. Our simulation studies and real data analysis confirm that MACQRM significantly outperforms MARM and conventional analyses of imaging data.

The Fallback And Multi-Stage Fallback Closed Group Procedures

✦Kenneth Liu, Merck & Co., Inc., , *Kenneth_Liu@Merck.com*; Duane Snavely, Merck & Co, Inc.

Key Words: Multiple endpoints, Multiple comparisons, Multiplicity, Closed testing, Strong control Type I error rate, Clinical trials

The "fallback closed group procedure" extends the fallback test, which works on one hypothesis at a time, to disjoint groups of hypotheses. First, ordered groups each receive part of the Type I error (a). If all hypotheses within a group are rejected using any closed testing procedure, then a accumulates, making tests of later groups more powerful. The "fallback closed group procedure" strongly controls the Type I error (a), uses any (possibly different) closed testing procedure, can be applied repeatedly to handle multiplicity problems of any dimension, and has a stepwise representation. Similarly, the "multi-stage fallback closed group procedure" extends the multi-stage fallback test and is also presented.

Net Effect Plots For Non-Elliptically Distributed Predictors

✦Xin Zhang, University of Minnesota, Twin Cities, 313 Ford Hall, 224 Church Street SE, Minneapolis, MN 55455 United States, *zhan0648@umn.edu*; R. Dennis Cook, University of Minnesota

Key Words: Net Effect Plots, Sufficient Dimension Reduction, Central Solution Subspaces

The net effect plot (Cook 1995) is a very useful tool for studying the contribution of selected predictors to a regression problem, with or without a pre-specified parametric model. We focus on graphical methods for studying the contribution of a subset of predictors after fully accounting for the contributions of the rest of predictors to the regression. While marginal plots might be misleading and added variable plots usually overestimate the importance of the selected predictors, net effect plots can reveal their true contribution. At the same time, sufficient dimension reduction in regression is a useful precursor that can facilitate the study. Specifically, using dimension reduction methods to find a sufficient distributional index function (Cook 1995) makes brushing, linking and analyzing net effect plots much easier. In particular, the methods based on Central Solution Subspaces (Li and Dong 2009) could be used even when the predictors are non-elliptically distributed. References: Cook, R. D. (1995). Graphics for studying the net e?ects of regression predictors. Li, B. and Dong, Y. (2009). Dimension Reduction for Nonelliptically distributed predictors.

Detection Of Central Dimension-Reduction Subspaces In Regression

✦SANTIAGO VELILLA, UNIVERSIDAD CARLOS III DE MADRID, DEPARTAMENTO DE ESTADÓSTICA, C/ MADRID, 126, GETAFE (MADRID), International 28903 SPAIN, *santiago.velilla@uc3m.es*

Key Words: Dimension reduction, Graphical regression, SIR and SAVE

Dimension reduction is a widely applied technique in regression. The basic problem in this field is the description of the central subspace (Cook, 1998), a linear manifold that helps to describe parsimoniously how the conditional distribution of a response variable changes with the values of a set of predictors. However, methods for searching directions inside the central subspace concentrate typically on a portion of it, imposing at the same time some assumptions on the marginal distribution of the regressors. Proposals for an exhaustive characterization of the central subspace exist, but they still depend on restricting the distribution of the regressors. This comunication presents a method for fully recovering the central subspace that places no restrictions on the predictors, other than the existence of first and second order moments. A data example is analyzed.

Diagnostic Tools For Hierarchical Linear Models

✦Adam M Loy, Iowa State University, 2414 Snedecor Hall, Ames, IA 50011, *aloy@iastate.edu*

Key Words: diagnostics, influential points, hierarchical linear models, R packages

Numerous familiar diagnostic tools exist for the linear regression model that enable us to check model assumptions and identify influential observations that might distort parameter estimates, predictions, and the precision of both. Less known are diagnostic measures for hierarchical linear models. Hierarchical linear models do not assume independence between data points, violating the usual modeling assumptions. Observations can be influential at multiple levels of a model. In the R package HLMdiag we have implemented residual analysis and case-deletion diagnostics for checking the model assumptions and detecting influential points in the hierarchical linear model. In particular we will emphasize graphical techniques for model checking and the detection and investigation of influential observations or groups of observations. We will present the functionality of these procedures using case studies.

Selecting The Optimal Window Size For Spatial Scan Statistics

✦Junhee Han, University of Arkansas, SCEN 315, University of Arkansas, Fayetteville, AR 72701, *falllunar@gmail.com*; Li Zhu, National Cancer Institute; Eric Feuer, National Cancer Institute; David Stinchcomb, National Cancer Institute; Zaria Tatalovich, National Cancer Institute

Key Words: Scan statistics, Windows Size, Cancer mortality, Disease Surveillance, SaTScan

The scan statistics is widely used in spatial, temporal, and spatio-temporal disease surveillance to identify areas of elevated risk and to generate hypotheses about disease etiology. In such a statistics, the area of the scanning window is allowed to vary which may take any predefined shape. It is very useful when we lack a prior knowledge about the size of the area covered by the cluster. But varying window shapes and sizes may produce different clustering patterns for the same data. This talk proposes a cluster information criterion that takes into account of likelihood, number of parameters, and power and size to evaluate the choices of varying window sizes. Simulation studies and real cancer incidence and mortality data show that the proposed cluster information criterion can identify the optimal window sizes for the purpose of disease surveillance.

311 Real World Applications of Statistical Learning and Data Mining Methods ■

Section on Statistical Learning and Data Mining
Tuesday, August 2, 8:30 a.m.–10:20 p.m.

Covariance Estimation And Variable Selection For High-Dimensional Psychiatric Data

✦Vivian H Shih, UCLA Department of Biostatistics, 21962 Yellowstone Lane, Lake Forest, CA 92630, *vivianhshih@gmail.com*; Catherine Ann Sugar, University of California, Los Angeles

Key Words: covariance estimation, variable selection, high dimensionality, neurocognition, pediatric psychiatry

Clinical investigators now routinely collect data on a large number of measures relative to their sample size. This is particularly true in psychiatry where test batteries include a multiplicity of data types such as clinical symptoms, behaviors, and neurocognitive performance, as well as measures derived from EEG, MRI and fMRI. It is of particular interest to identify phenotypes or patterns of deficits characterizing specific disorders. CIDAR: Translational Research to Enhance Cognitive Control is a study examining cognitive control deficits in children with ADHD and Tourettes. We explore patterns of baseline characteristics based on a restrictive subset of ~500 variables for 367 subjects. In this preliminary analysis, we apply variants of cutting-edge tools (e.g. sparse covariance estimation, graphical/network models, variable selection methods based on adaptive shrinkage), tailored to incorporate the known superstructure of our data. We use these modified techniques to look for differences in the relational structure of our measures across groups and to identify predictors of behavioral symptoms and academic performance for use in subsequent longitudinal analyses.

A Poisson Regression Examination Of The Relationship Between Website Traffic And Search Engine Queries

✦Heather L R Tierney, College of Charleston, Dept of Econ and Finc, 5 Liberty Street, Charleston, SC 29401, *hlrtierney@yahoo.com*; Bing Pan, College of Charleston

Key Words: Poisson Regression, Search Engine, Google Insights, Aggregation, Normalization Effects, Scaling Effects

A new area of research involves the use of normalized and scaled Google search volume data to predict economic activity. This new source of data holds both many advantages as well as disadvantages. Daily and weekly data are employed to show the effect of aggregation in Google data, which can lead to contradictory findings. In this paper, Poisson regressions are used to explore the relationship between the online traffic to a specific website and the search volumes for certain keyword search queries, along with the rankings of that website for those queries. The purpose of this paper is to point out the benefits and the pitfalls of a potential new source of data that lacks transparency in regards to the original level data, which is due to the normalization and scaling procedures utilized by Google.

Assessing The Repeatability Of Functional Data: Repeatability Of Tissue Fluorescence Measurements For The Detection Of Cervical Intraepithelial Neoplasia

✦Jose-Miguel Yamal, University of Texas School of Public Health, 1200 Herman Pressler, RAS W928, Houston, TX 77030, *Jose-Miguel.Yamal@uth.tmc.edu*; E. Neely Atkinson, University of Texas M.D. Anderson Cancer Center; Dennis Cox, Rice University; Michele Follen, Drexel University College of Medicine

Key Words: functional data, repeatability, cervical cancer, classification, linear mixed-effects

We examined the differences in 378 repeated spectroscopic measures of the cervix. To assess the repeatability of this functional data, we deconstructed the differences into a shape difference measure and an intensity difference measure. We then examined causes of variability

and the importance of the order of measurements. Finally, we examined the classification concordance of cervical intraepithelial neoplasia between the repeat measurements.

Local Frequency Based Estimators For Anomaly Detection In Oil And Gas Applications

✦Choudur K Lakshminarayan, Hewlett Packard laboratories, 142321 Tandem Boulevard, Austin, TX 78728 USA, *choudur.lakshminarayan@hp.com*; Evan Kriminger, University of Florida, Gainesville; AlexanderSingh Alvarado, University of Florida, Gainesville

Key Words: Fourier analysis, Dynamical systems, turbulence, Oscillations, Non-linear tiem series, frequency methods

Modern industrial applications such as the smart grid and oil and gas are continuously monitored. The massive amounts of data collected is then processed, and analyzed to generate actions to ensure smooth operations to positively impact the bottom line. In the oil and gas industry, modern oil rigs are outfitted with thousands of sensors to measure the flow rates, as well as the physical and chemical characteristics that affect production from underground off-shore and on-shore reservoirs. Analytical methods packaged into a surveillance system and applied to the massive network of sensors track the state of the system and issue warning alerts about impending failures. In this setting, real time algorithms are needed to detect a diversity of event types, such as anomalies, trends or forewarn failure events to generate alerts for proactive engineering actions. In this paper, online algorithms are applied to quickly detect anomalies,and turbulence in the flow of oil in the bore well which is typical in oil production. The short time Fourier transform and dynamical systems were utilized to uncover structure in the data.

Authorship Discrimination And Topic Modeling: The Federalist Papers

✦Mario Andres Morales, Hunter College and Polytechnic Institute of NYU, Dept of Chemical and Biological Sciences Bioinformatics, 6 MetroTech Place, Brooklyn, NY 11201, *mmo0010@hunter.cuny.edu*

Key Words: Text Mining, Topic Modelling, Latent Dirichlet Allocation, Federalist Papers

After forty seven years since the publication of the seminal work of Mosteller and Wallace about the use of Bayesian reasoning to assign the authorship to the disputed federalist papers, many other approaches have been used to replicate similar results based on the features described in this analysis. In this paper we reviewed the authorship problem, we cleaned the Federalist corpus with the use of desktop tools for natural language processing with python and the statistical programming language R and for the first time we estimated a topic model using the Latent Dirichlet Allocation model of Blei, et. al with the goal of differentiating authorship based on the estimated topics.

Spatio-Temporal Models For Wireless Network Data

✦Ganesh K Subramaniam, AT&T Labs, 180 Park Avenue, Florham Park, NJ 07932 USA, *mkg@research.att.com*; Colin Goodall, AT&T Labs; Ravi Varadhan, Division of Geriatrics & Gerontology, Johns Hopkins School of Medicine

Key Words: spatio-temoral, time series, mobility network, functional data analysis

Spatial-temporal models arise when data are collected across both space and time. With AT&T network data, a typical example would be that of a monitoring data on the mobility network (a network of towers) on which data are collected at regular intervals, say on a monthly basis. We have a time series associated with usage of minutes (voice) and Kb (data) for every tower located throughout the country. Thus the analysis has to take account of spatial dependence among the towers, but also that the observations at each tower typically are not independent but form a time series. In other words, one must take account of temporal correlations as well as spatial correlations. The topic of interest is how do the temporal patterns associated with the time series of a given tower correlate to temporal patterns in neighboring towers. We use a sample of time series from the network data to explore this question.

312 R Programming

Section for Statistical Programmers and Analysts, Section on Statistical Computing, Section on Statistical Graphics

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Rexcel For Pharmaceutical Applications

✦Richard Heiberger, Temple University Department of Statistics, 332 Speakman Hall 006-12, 1810 N. 13 St., Philadelphia, PA 19122-6083, *rmh@temple.edu*

We give examples of an Excel frontend and an R backend for designing complex processes and analyses, with the data and control coming from the spreadsheet environment.

R Commander: A Graphical User Interface For R

✦Jason Wilson, Biola University, 13800 Biola Ave, La Mirada, CA 90639, *jason.wilson@biola.edu*

Key Words: R, Education, R Commander, Rcmdr, Introduction to Statistics

R Commander is one of the most popular graphical user interfaces (GUI) for R. Some of the reasons for its popularity are that it is: (1) easy to use, (2) actively maintained, (3) user extensible, and (4) incorporated into RExcel (program implementing R from within the Excel framework). In addition, it helps bridge users to the Command Window by displaying the R code it produces in a separate window. This talk will introduce the R Commander package, its capabilities, and comment on its use in an Introduction to Statistics classroom.

Constructing Tabular Output Using The Graphics Device (The R Package Tabular)

✦Carlin Brickner, The Visiting Nurse Service of New York, 1250 Broadway - 20th Floor, VNSNY: Research, New York, NY 10001, *carlin.brickner@vnsny.org*; Rocco Napoli, The Visiting Nurse Service of New York

Key Words: R, Tabular Output, Reproducible Research

Statisticians are often called upon to produce tabular output for papers and presentations. Clients expect a certain quality of output that does not match the default tabular output from statistical software. Various time-consuming, non-reproducible, error prone or manual methods (copy/paste/export to spreadsheet applications, custom programming, etc.) are often used to "dress up" the output. The R package 'tabulaR' is a proposed solution to this problem. This package utilizes the object oriented language, and assumptions about the structure of tabular data to drive the presentation of a table completely within the R environment. "Dress up" features are accessible through a user friendly interface familiar to the R programmer, as well as all of the file formats available for export via the R graphics device. These features include: text markup, formatting to grouped row and column label hierarchies, column and row dividers, conditional highlighting, footnotes, footer/header, and page overflow management.

Density Estimation Packages In R

✦Henry Deng, Rice University, 1605 Rice Blvd, Houston, TX 77005 USA, *hd4@rice.edu*

Density estimation is an important statistical tool, and within R, there are over 10 packages for density estimation, including np, locfit, and KernSmooth. At different times, it is often difficult to know which to use. In this project, we will summarize the results of our study comparing these packages. We will present a brief outline of the theory behind each package, as well as a description of the functionality and comparison of performance. Some of the factors we touch on are dimensionality, flexibility, and control over bounds.

Clinical Trials Tables And Listings Using R

✦Young Kim, Pharmanet, Inc., 1000 CentreGreen Way, Suite 300, Cary, NC 27513, *ykim@pharmanet.com*

Key Words: R, LaTeX, Sweave, TLF, Tables, Listings

In the pharmaceutical industry, biostatisticians and statistical programmers plan and create summary tables, listings, and figures of clinical trial data for the inclusion in clinical study reports. With the availability of R and other freeware, it is now possible to create high quality tables and listings specified in most statistical analysis plans. In this talk, methods for preparing tables and listings are presented along with examples. Future work is also discussed to improve the process of creating tables and listings.

A Seasonal-Trend Decomposition Procedure Using Kz-Filters

✦Brian Close, SUNY-Albany, Albany, NY 12440, *brian.close@gmail.com*; Igor Zurbenko, State University of New York at Albany

Key Words: seasonal adjustment, time series, changepoint, kza, kz, kzft

KZ is a filtering procedure that allows decomposing a time series into trend, seasonal, and remainder components. KZ has a simple design that consists of a sequence of applications of the KZ low pass filter; the simplicity allows analysis of the properties of the procedure and allows fast computation, even for a long time series and large amounts of trend and seasonal smoothing. Other features of KZ are specification of different seasonal periods and trend smoothing that range, in a nearly continuous manner; robust estimates of the trend and seasonal components that are not distorted by aberrant behavior in the data (adaptive filtration); nonparametric specification of periods of various seasonal components; and the ability to decompose a time series with missing values or any longitudinal data. An example of data from fatal accidents on US roads will be discussed, necessary software is provided in R.

313 Spatial and Spatio-Temporal Modelling

Section on Statistics and the Environment

Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Compactly Supported Multivariate Covariance Matrix Functions

✦Juan Du, Department of Statistics, Kansas State University, 108D Dickens Hall, Department of Statistics, Manhattan, KS 66506, *dujuan@ksu.edu*; Chunsheng Ma, Department of Mathematics and Statistics, Wichita State University

Key Words: Covariance matrix function, Covariance tapering, Cross covariance, Direct covariance, Multivariate random field, Variogram matrix function

Covariance tapering is a useful technique to mitigate the numerical burdens in dealing with the large spatial data sets. This technique is applied to multivariate case and compactly supported multivariate covariance functions are needed for multivariate tapering functions. To meet this need, we construct a class of multivariate random fields in R^d whose direct and cross covariance functions are compactly supported by using the convolution approach. In addition, a class of second-order stochastic processes whose direct and cross covariance functions are of PÛlya type is also derived. Simulation study is conducted to show the computational gain and application in cokrigging by using proposed multivariate tapering function.

Transformed Gaussian Markov Random Fields: A New Generalized Class Of Random Fields To Incorporate Spatial Dependence

✦Marcos Oliveira Prates, University of Connecticut, 348C Foster Dr, Willimantic, CT 06226, *marcosop@yahoo.com*; Dipak K Dey, Univerisity of Connecticut; Michael R Willig, University of Connecticut; Jun Yan, University of Connecticut

Key Words: Bayesian modeling, Bayesian network, GLMM, GMRF, undirected graphical model

Gaussian random field (GRF) and Gaussian Markov random field (GMRF) have been widely used to accommodate spatial dependence. For a variety of datasets the use of the Generalized Linear Mixed Models (GLMM) framework is most appropriate (e.g., Count data). Although GRFs and GMRFs have been widely used under a GLMM setup to accommodate spatial dependence, this method presents some drawbacks due to the inherent symmetry and weak tail of the Gaussian distribution. In this paper, a new class of random fields, denominated transformed Gaussian random field (TGRF) and a new class of Markov random fields, called transformed Gaussian Markov random field (TGMRF), are introduced. Both TGRF and TGMRF are constructed

performing marginal transformations under GRF and GMRF respectively. Due to the applied transformation it allows a general and flexible representation that can easily accommodate both asymmetry and heavy tail when necessary. Besides that, the TGRF mimics the GRF with respect to several of its properties. We illustrate the importance of the developed methodology with a simulation and an ecological study.

A Generalization Of The Neyman-Scott Process

✦Chun Yip Yau, Chinese University of Hong Kong, Hong Kong, *cyyau@sta.cuhk.edu.hk*; Ji Meng Loh, AT&T Labs-Research

Key Words: Neyman-Scott process, K-function, Gibbs process, Regular point process

In this paper we introduce a generalization of Neyman-Scott process that allows for regularity in the parent process. In particular, the parent process is a Strauss process, and the offspring process is uniform on a disc centered at each parent. Such a generalization allows for point realizations that show a mix of regularity and clustering in the points. We work out a closed form expression of the K function for this model and use this to fit the model to data. The approach is illustrated by applications to the locations of a species of trees in a rainforest dataset.

Spectral Analysis Of Spatio-Temporal Processes On The Sphere

✦Chunfeng Huang, Indiana University, Statistics House, 309 North Park Ave, Bloomington, IN 47408 USA, *huang48@indiana.edu*; Yoon-Jin Lee, Indiana University; Scott Robeson, Indiana University; Haimeng Zhang, Mississippi State University

Key Words: Spatio-Temporal, Sphere, Spectral, homogeneous, stationary

Geophysical and environmental processes generally vary in both space and time, and their spatial domain is the spherical Earth. When the process is homogeneous on the sphere and stationary with respect to the time, a spectral representation of the spatio-temporal covariance function on the sphere is presented. Several family of valid covariance functions are provided to model the spatio-temporal processes on the sphere.

Nonparametric Covariance Function Estimation In Isotropic Spatial Process

✦Yang Li, Iowa State University, 2505 Aspen Rd Unit 3, Ames, IA 50010, *yangli@iastate.edu*; Zhengyuan Zhu, Iowa State University

Key Words: covariance function, covariogram, nonparametric estimation, isotropic process

In spatial statistics estimating covariance structure is of fundamental importance. In this paper, we analyze a nonparametric method of estimating the covariance function for isotropic process. By Bochner's theorem, covariance function can be approximated as a discrete summation of oscillating Bessel functions with properly chosen nodes. Genton and Gorsich (2002) among others have found that using roots of Bessel functions as nodes has theoretical advantages over the nodes chosen by ad hoc method. Based on these findings we develop new methods for nonparametric covariance estimation which is efficient and easy to implement, with the number of nodes decided by a model selection method based on BIC to reduce over-fitting problems. Comparison with the well-known kernel smoothed estimation method is also performed in simulation studies.

Bayesian Estimation Of Multivariate Matern Covariance Parameters

✦Bonnie B. Terry, Baylor University, 101 North 1st street, Cranfills Gap, TX 76637, *bonnie_terry@baylor.edu*; Jane Harvill, Baylor University

Key Words: Multivariate geostatistics, Matern covariance function, Bayesian estimation

Covariance modeling and estimation is key to spatial prediction methods. Univariate techniques have been widely studied and implemented. Often, however, significant gains may be realized from simultaneously modeling the spatial dependance structure of two or more attributes. Gneiting, et al. (2010) develop a valid Matern cross-covariance function for multivariate Gaussian random fields. We employ Bayesian methodology to parameter estimation for the multivariate Matern model.

314 Methodology for Discrete Data

Section on Statistics in Epidemiology, Biometrics Section
Tuesday, August 2, 8:30 a.m.–10:20 a.m.

Estimation Of Sample Size For Case-Control Studies By Decomposition Of

✦Peng Tu Liu, FDA, 5100 Paint Branch Pkwy, College Park, MD 20740, *peng.liu@fda.hhs.gov*

Key Words: Obserable Constraints, Minimum Detection, Continuous Exposure, Optimum Strategy, Programming Technique, Retrospective Study

In a case-control study, when the exposure measurement is a dichotomous variable, then the population is distributed as a mixture of two Bernoulli distributions. When the exposure measurement is a continuous variable, such as a log-normal distribution, then the population is distributed as a mixture of two continuous distributions. Suppose the disease rate and the population exposure distribution are known and the minimum odds ratio (or relative risk) we desire to detect is selected. Then we can decompose a known population exposure distribution into two disease-specific exposure distributions to satisfy the sample size equation for testing two independent proportions or two independent scores.

A Comparison of Predictive Marginals Estimated from Logistic Regression Models and Log-Linear Regression Models with Categorical Data

✦Chaoyang Li, Centers for Disease Control and Prevention, 1600 Clifton Road NE, MS E97, Atlanta, GA 30333 USA, *cli@cdc.gov*; Earl S Ford, Centers for Disease Control and Prevention; Catherine A Okoro, Centers for Disease Control and Prevention; Tara W Strine, Centers for Disease Control and Prevention; Jin-Mann S Lin, Centers for Disease Control and Prevention; Lina S Balluz,

Centers for Disease Control and Prevention

Key Words: predictive marginals, prevalence, complex survey data, standardization, diabetes, obesity

Predictive marginals have been used as a useful tool to compare adjusted prevalence estimates between subgroups controlling for differences in the distribution of covariates in survey data analyses. Little is known about whether different regression models may yield similar predictive marginal estimates. We analyzed data from the 2008 Behavioral Risk Factor Surveillance System to empirically compare the predictive marginal estimates of obesity, diabetes, and myocardial infarction among non-Hispanic whites, non-Hispanic blacks, Hispanics, and adults with other race/ethnicity using logistic regression and log-linear regression analyses. Results showed that with adjustment for age and sex, log-linear regression models yielded predictive marginal estimates similar to the age- and sex-adjusted prevalence estimates obtained from the direct standardization, whereas logistic regression models yielded inflated predictive marginal estimates (relative difference ranged from 29.7% to 75.2%). Log-linear regression models appear to perform better than logistic regression models for estimating predictive marginals, particularly when the population prevalence is low and/or sample size is small.

Two Artificial Mixture Methods For Discrete/ Grouped Failure Time Data

✦Shufang Wang, University of Michigan, 1420 washington heights, Department of Biostatistics, Ann Arbor, AL 48109 US, *sfwang@umich.edu*; Alexander Tsodikov, University of Michigan

Key Words: discrete failure time, artificial mixture model, proportional odds model

we consider a general discrete transformation model for failure time data in a large data set with many ties by changing the model form at the "complete-data" level (conditional on artificial variables). Two complete data representations of a given discrete transformation model are studied: proportional hazards (PH) and proportional odds (PO) mixture methods. In PH mixture method, we reduce the high-dimensional optimization problem to many one-dimensional problems. In PO mixture method, a recursive procedure is available to simplify the optimization. As a result, we advocate the PO mixture method.

Compare Predicted Counts Between Groups Of Zero Truncated Poisson Regression Model Based On Recycled Predictions Method

✦Yan Wang, UCLA School of Public Health, Department of Biostatistics, , *wangyan@ucla.edu*; Michael Ong, University of California, Los Angeles; Honghu Liu, UCLA School of Dentistry

Key Words: Zero Truncated Poisson (ZTP) regression model, recycled predictions method, variance estimation

Zero Truncated Poisson (ZTP) regression model is used to model positive count data, where zero is a potential value but is almost impossible to be observed due to the nature of study and its design. ZTP is more accurate than traditional Poisson regression model for this kind of data. In practice, researchers often need to test the difference of the predicted counts between groups with ZTP regression model. The test result can be misleading if the design is very unbalanced. However, the combination of ZTP regression model and recycled predictions method is one possible way to create an identical structure of the covariates when comparing the predicted counts between groups. This paper uses ZTP regression model based on recycled predictions method to model the positive count data and estimates the variance of the difference of the predicted counts by delta method. Finally, the model and estimation techniques are applied to a real study of Adherence and Efficacy of Protease Inhibitor Therapy (ADEPT).

Modeling Zero-Inflated Continuous Data with Varying Dispersion

✦Ka Yui Karl Wu, The University of Hong Kong, Department of Statistics & Actuarial Science, Meng Wah Complex, Hong Kong, International China, *karlwu@hku.hk*; Wai Keung Li, University of Hong Kong

Key Words: EM Algorithm, Generalized Linear Model, Overdispersion

Zero-inflated data are often observed in empirical studies of different scientific fields. Data are considered as zero-inflated if the observed values of a random vector contain significantly more zeros than expected. Excessive occurred zeros to the dependent variable in a regression model discourage straightforward modelling by classical regression techniques. In the past, zero-inflation is considered as a count data problem and Zero-Inflated Poisson regression (ZIP) has been established to be the standard tool for zero-inflation modelling. The approach is based on a joint probability density function in which the probability for non-zero observations and response mean are both parameters and interlinked by two pseudo-simultaneously estimated linear models. However, constant dispersion is often assumed even when overdispersion is a common feature in almost every empirical data set. In our paper, the dispersion is formulated as a gamma generalized submodel interlinked with a mean and a zero-inflation probability submodel. We propose a modified triple, nested iterative approach to model response mean, dispersion and zero-inflation probability simultaneously.

In Defence Of An Ecologic Study Esign Using Hospital Discharge Data

✦Lawrence Lessner, Institute of Health and the Environment, State University of New York, Albany, 26 Wilan Lane, Albany, NY 12203, *LLessner@nycap.rr.com*

Key Words: Ecologic study, environmental health, confounding, contextual study

In Defense of an Ecologic Study Design Using Hospital Discharge Data Lawrence Lessner Institute of Health and the Environment State University of New York, Albany 26 Wilan Lane, Albany, NY 12203 LLessner@nycap.rr.com An ecological study (abbreviated ES) here is a study where the unit of analysis is a group of people rather then an individual level study (abbreviated ILS) whose unit of analysis is a person. Frequently the objective in an ES is an individual level parameter such as prevalence, incidence, or mortality. There is a considerable body of epidemiological literature that has identified a number of serious problems with using an ES to estimate ILS parameters. Our objective is to present an ecological study design using hospital discharge data for a number of health outcomes, and to assess the validity of this approach using the criticisms from S. Greenland, Ecologic Versus Individual-level sources of Bias in Ecologic Estimates of Contextual

Health Effects, 2001 IJE, 30, 1343-1350. The study design using hospital discharge data is a very flexible and epidemiological useful study design.

Analysis Of Progression Free Survival Data Using A Discrete Time Survival Model That Incorporates Measurements With And Without Diagnostic Error

✦Sally Hunsberger, The National Cancer Institute, 6130 Executive Blvd, Bethesda, MD 20892 USA, *sallyh@ctep.nci.nih.gov*; Albert Paul, NICHD; Lori Dodd, NIAD

Key Words: Gaussian random effects, Sensitivity, Specificity, Conditional Independence

In cancer studies Progression Free Survival is an endpoint that is becoming very important in the development of new theraputic agents. Two methods of determining Progression are typically used: 1) the local radiologist evaluates scans and 2) scans are reviewed by and independent (central) reviewer. The second method is considered to be a gold standard (GS) but is expensive, time consuming and logistically difficult. The first method has measurement error associated with it but is less expensive and easier to obtain. When PFS data using the test with measurement error are analyzed, inferences about covariate effects may be invalid. A sampling strategy is evaluated where data are collected on a subset of subjects using the GS test and on all subjects using the test that has error. The strategy is designed to maintain valid inferences while requiring the more expensive or difficult test on a small proportion of patients. We explore the effect of different diagnostic test properties on inference via simulation and use the methodology to analyze a renal cancer example.

315 ASA College Stat Bowl II

ASA, ENAR, WNAR, IMS, SSC, International Chinese Statistical Association, International Indian Statistical Association

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

316 Complex and massive multiple testing problems ■●

General Methodology, Biopharmaceutical Section, ENAR, International Chinese Statistical Association, Section on Health Policy Statistics

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Multiple Testing with Complex Hierarchical Structures in Clinical Trials

✦Alex Dmitrienko, Eli Lilly and Company, 10913 West 144th St, Overland Park, KS 66221, *alexei@lilly.com*; Ajit C Tamhane, Northwestern University

Key Words: Multiple comparisons, Familywise error rate, Gatekeeping procedures, Clinical trials

This talk discusses approaches to addressing multiplicity issues in clinical trials with multiple objectives. The objectives define multiple testing problems with complex hierarchical structures, eg, problems with null hypotheses grouped into multiple families with general logical relationships. A new method has been developed for defining multiple testing procedures for problems with multiple families of null hypotheses (termed gatekeeping procedures). The method is based on constructing mixtures of procedures used in individual families, including p-value-based and parametric procedures, and is termed the mixture-based method. Resulting gatekeeping procedures account for arbitrary logical relationships and control the familywise error rate in the strong sense. This general method is illustrated using clinical trials with multiple objectives, including multiple dose-placebo comparisons/patient populations and multiple dose-placebo comparisons/noninferiority-superiority tests.

Practical Implementations of Decision-Theoretic Multiple Inferences

✦Peter Westfall, Texas Tech University, Department of Information Systems and Quantitative, Lubbock, TX 79409-2101, *peter.westfall@ttu.edu*; Russ Wolfinger, SAS Institute Inc; Randy Tobias, SAS Institute; Ananda Managa, Sam Houston State University

Key Words: Multiple Comparisons, Multiple Testing, Familywise Error Rate, False Discovery Rate, Bayesian Methods, Scale up

With so many multiple comparisons methods to choose from, practitioners face a bewildering array of choices. And they just want to know, "which is the best method?" Decision theoretic methods offer an answer to the question; and modern computing allows simple, Monte Carlo-based solutions. The problem is then determining the loss functions. The question of "scientific loss" due to Type I and Type II errors is considered carefully in the context of genomic data, loss functions are carefully developed, and recommendations are given. The methods are implemented using SAS software.

Permutation Multiple Tests of Binary Features May Not Control Error Rates

✦Eloise Kaizar, Ohio State University, OH 43210, *ekaizar@stat.osu.edu*; Yan Li, Amylin Pharmaceuticals; Jason C Hsu, Ohio State University

Key Words: Multiple tests, Permutation, FWER, pharmacogenomics

Multiple testing for significant association between predictors and responses has a wide array of applications. One such application is pharmacogenomics, where testing for association between responses and many genetic markers is of interest. Permuting response group labels to generate a reference distribution is often thought of as a convenient thresholding technique that automatically captures dependence in the data. In reality, non-trivial model assumptions are required for permutation testing to control multiple testing error rates. When binary predictors (such as genetic markers) are individually tested by standard tests, permutation multiple testing can give incorrect unconditional and, especially, conditional assessment of significances, and thus misleading results.

317 Current Advances in Modeling Time Series of Counts, with Applications ■●

Business and Economic Statistics Section, International Chinese Statistical Association

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Autoregressive Models for Count Time Series

✦Konstantinos Fokianos, University of Cyprus, Math and Stat Dept, PO Box 20537, Nicosia, 1678 Cyprus, *fokianos@ucy.ac.cy*

Key Words: estimation, prediction, stationarity, autocorrelation function, covariates, modeling

Count time series are observed in diverse applications, for instance consider the number of transaction per minute of some stock, or the monthly number of people with a certain disease, and so on. For the analysis of these data, there has been developed a number of models based either on thinning operator or on GLM framework. We will be examining the second class of models which include a feedback mechanism. Such models are expected, in general, to be more parsimonious, pretty much as is the case of GARCH models. It is important therefore to study their statistical properties and develop algorithms for estimation. In particular we will be discussing some multivariate generalizations of previously established models in the literature.

A Flexible Hierarchical Approach to Modeling Discrete-Valued Spatio-Temporal Data

✦Scott H Holan, University of Missouri, Department of Statistics, 146 Middlebush Hall, Columbia, MO 65211, *holans@missouri.edu*; Christopher K. Wikle, University of Missouri

Key Words: Bayesian hierarchical models, counts, overdispersion, spatially-varying

In many cases modeling discrete-valued spatio-temporal data is a straightforward endeavor. However, in many real-world applications the complexities of the data and/or process don't allow for routine model specification. For example, often discrete-valued spatio-temporal data exhibit zero-inflation, over/under dispersion or heavy tails and contain many sources of uncertainty. In order to accommodate such structure, while quantifying different sources of uncertainty, we propose a hierarchical Bayesian model that utilizes a flexible likelihood specification. The approach we propose allows the likelihood to adapt to the nuances of the discrete-valued data while flexibly accommodating different spatio-temporal dependence structures. The effectiveness of our methodology is demonstrated through simulation and through a real-data application.

Forecasting Periodic Discrete-Valued Time Series

✦David S. Matteson, Cornell University, 282 Rhodes Hall, Ithaca, NY 14853 United States, *dm484@cornell.edu*

Key Words: Dynamic factor model, Smoothing splines

We introduce a new method for forecasting that combines discrete-valued time series models with a dynamic latent factor structure. The factor structure models the observed non-stationary patterns in periodic data and greatly reduces the number of model parameters. The factor model is combined with stationary discrete-valued time series models to capture the remaining serial dependence in the intensity process. We compare frequentist and Bayesian estimation methods to forecast and conduct inference for applications in staffing and manufacturing.

Stochastic Models for Multivariate Time Series of Counts, with a Marketing Application

✦Nalini Ravishanker, University of Connecticut, Department of Statistics, U-4120, 215 Glenbrook Road, Storrs, CT 06269-4120 USA, *nalini.ravishanker@uconn.edu*; Rajkumar Venkatesan, University of Virginia; Shan Hu, University of Connecticut

Key Words: Attitudes, Bayesian modeling, Discrete-valued time series, Pharmaceutical marketing data

In several applications, there is an increasing need for accurate modeling of multivariate time series of counts for several subjects as functions of relevant covariates (subject-specific and time-varying), incorporating dependence over time and dependence between the components of the response vector. This talk describes a hierarchical dynamic non-linear model framework with a marketing application, using data from a multinational pharmaceutical firm. We model multivariate count data responses on the number of monthly prescriptions made by physicians for the focal drug from this firm, and for drugs from its competitors. We incorporate into the analysis customer attitudes from a survey obtained at irregular times, and discuss imputation of missing attitudes and customer lifetime value computation.

318 Outcome Dependent Sampling Designs for Correlated Longitudinal Data ■●

ENAR, International Chinese Statistical Association, Section for Statistical Programmers and Analysts, Section on Health Policy Statistics, Section on Statistics in Epidemiology

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Semiparametric Estimation for Longitudinal Binary Response Model Parameters Under Outcome-Dependent Sampling Designs

✦JONATHAN SCHILDCROUT, Vanderbilt University School of Medicine, 1161 21st Ave South, S-2323 Medical Center North, NASHVILLE, TN 37206 United States, *jonathan.schildcrout@vanderbilt.edu*

Key Words: longitudinal data, outcome dependent sampling, epidemiological study designs

We will discuss a general semi-parametric estimation strategy for analyses of longitudinal binary response data ascertained from outcome dependent sampling designs. The strategy permits time-varying and time-invariant estimation targets. It can be used for prospective or retrospective designs where sampling is based on an ancillary variable that is related but not equal to the response vector. The strategy involves two models. The first (sampling variable) model is estimated with a logistic regression of the sampling variable on the exposure and response variables. Estimates from this model are then used in an offset term for the

second (target) model that regresses the response variable on exposure. The offset term is used to acknowledge or correct for the biased study design. The second model, in particular, is estimated with covariance weighted generalized estimating equations. We will discuss several designs and will explore optimal designs for various estimation targets.

Subject-Specific Analysis of Longitudinal Binary Data from Outcome-Dependent Sampling Designs

✦John Neuhaus, University of California, San Francisco, Division of Biostatistics, 185 Berry Street, Lobby 5, San Francisco, CA 94107-1762, *john@biostat.ucsf.edu*; Alastair Scott, University of Auckland; Chris J. Wild, University of Auckland; Charles McCulloch, U of California, San Francisco

Key Words: Mixed-effects models, Profile likelihood, Retrospective sampling

Investigators often use case-control or, more generally, outcome dependent sampling designs to study rare binary outcomes. In a series of papers, Neuhaus, Scott and Wild developed a profile/pseudo-likelihood approach that addresses the clustering and differential sampling rates that such designs feature. Longitudinal follow up of subjects gathered in an initial outcome dependent sample can be used to study the trajectories of responses over time and to assess the association of changes in predictors within subjects with change in response. This talk shows that by augmenting the response of the original profile/pseudo-likelihood approach we can extend it to accommodate longitudinal data from a wide variety of complex and novel outcome dependent sampling designs. Data from a study of Attention Deficit Hyperactivity Disorder in children motivates the work and illustrates the findings.

Conditional Likelihood Approaches for Outcome Dependent Sampling Designs Using Continuous Longitudinal Data

✦Patrick Heagerty, university of washington, WA United States, *heagerty@uw.edu*; JONATHAN SCHILDCROUT, Vanderbilt University School of Medicine

Key Words: design, longitudinal

With the emergence of large-scale capture of patient data in electronic medical records there are opportunities to use these data bases to design novel biomedical investigations. When individual subjects have longitudinal measurements already recorded but interest is in linking change over time to new markers and/or to patient environmental characteristics then patients would need to be selected for additional data collection. We discuss sampling designs that generalize standard case-control designs to the longitudinal setting, and discuss ascertainment-corrected inference using longitudinal regression models conditional on selection into the study. We compare the efficiency of proposed designs to alternative random sampling and/or full cohort data collection.

Outcome and Probability Dependent Sampling Designs and Inference

✦Haibo Zhou, University of North Carolina, Chapel Hill, NC 27514 United States, *Zhou@bios.unc.edu*

Key Words: Outcome dependent sampling, probability dependent sampling

Biomedical studies are often designed to assess the relationship between some exposure X of interest and the corresponding outcome Y of individual adjusted by some confounding covariates Z. Restricted by the costs associated with exposure ascertainment, the full assessment of X on the whole study cohort is often not feasible. Two-stage stratified sampling design, introduced by Neyman (1938), is often used to enhance efficiency. At the first stage of a typical two-stage design, a relatively large random sample is drawn and measured for Y and Z, while, ascertainments of X are made at the second stage for a subsample drawn randomly, without replacement from the first stage data. Greater efficiency can be obtained through the two-stage sampling design (e.g. Breslow and Cain, 1988; Breslow et al., 2003 and Wang and Zhou, 2010). Another method for improving study efficiency is through biased sampling using the outcome-dependent-sampling (ODS) scheme. For example, the case-control study (e.g. Anderson, 1972; Prentice and Pyke, 1979) is the most well-known such design to deal with binary outcomes, and from it many subsequent designs have emerged. Among others, case-cohort studies were introduced by Prentice (1986) in order to reduce the cost by observing fewer subjects rather than following the whole cohort. Lu and Tsiatis (2006) propose a new way of estimating parameters in the linear transformation model component for the case-cohort study. Zheng et al (2010) describe likelihood-based approaches for the combining family-based and population-based case-control data. Schildcrout and Heagerty (2008) describe sampling based on the presence/absence of binary response series variation and propose conditional maximum-likelihood analyses. Biased sampling schemes can be a cost effective way to enhance study efficiency. In this paper, we propose a new two stage sampling design, the probability sampling scheme, in which, the second stage supplement samples are drawn based on a sampling probability calculated from the first stage data. The basic idea is to oversample those X that are on the two tails of its distribution. A semiparametric empirical likelihood inference procedure is proposed and the asymptotic normality properties of the proposed estimator is developed. Simulation results indicate that the sampling scheme and the proposed estimator is more efficient than the existing outcome dependent sampling design and the random sampling designs. We illustrate the proposed method with a data set from an environmental epidemiologic study, to assess the relationship between maternal polychlorinated biphenyl level and children's IQ test performance.

319 Bayesian Analysis Invited Session

Section on Bayesian Statistical Science, International Indian Statistical Association

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Hierarchical Bayesian Modeling of Hitting Performance in Baseball

✦Shane Jensen, The Wharton School, , *stjensen@wharton.upenn.edu*; Blakeley McShane, The Wharton School; Abraham Wyner, The Wharton School

Key Words: baseball, hidden Markov model, hierarchical Bayes

We have developed a sophisticated statistical model for predicting the hitting performance of Major League baseball players. The Bayesian paradigm provides a principled method for balancing past performance with crucial covariates, such as player age and position. We share information across time and across players by using mixture distributions to control shrinkage for improved accuracy. We compare the performance of our model to current sabermetric methods on a held-out season (2006), and discuss both successes and limitations.

Selection Sampling from Large Data Sets for Targeted Inference: Applications in Flow Cytometry

✦Ioanna Manolopoulou, Duke University, Box 90251, Duke University, Durham, NC 27708-0251, *im30@stat.duke.edu*; Cliburn Chan, Duke University; Mike West, Duke University

Key Words: flow cytometry, large data sets, mixture models, rare events, selection sampling, sequential

One of the challenges in using Markov chain Monte Carlo on very large datasets is the need to scan through the whole data at each iteration of the sampler. Here we consider the specific case where most of the data from a mixture model provide little or no information about the parameters of interest, and we aim to select subsamples such that the information extracted is most relevant. The motivating application arises in flow cytometry, where several measurements from a vast number of cells are available. Interest lies in identifying specific rare cell subtypes and characterizing them according to their corresponding markers. We present a Markov chain Monte Carlo approach where an initial subsample of the full dataset is used to guide selection sampling of a further set of observations targeted at a scientifically interesting, low probability region. We define a sequential strategy where the targeted subsample is augmented sequentially as estimates improve, and introduce a stopping rule for determining the size of the targeted subsample. An example from flow cytometry illustrates the ability of the approach to increase the resolution of inferences for rare cell subtypes.

Desiderata for a Predictive Theory of Statistics

✦Bertrand Salem Clarke, Dept. Medicine, Univ. Miami, 1120 NW 14th Street, CRB 611 (C-213), Miami, FL 33136, *bclarke2@med.miami.edu*

Key Words: prequentialism, complexity, variance-bias, stability, predictor updating

We present a unified treatment for how to approach predictive problems. It is based on six `desiderata' which, taken together, are an effort to clarify what criteria a good predictive theory of statistics should satisfy. The motivation for this work is that there are many contexts where predictive validation is more important than model identification, which may be practically impossible. This is particularly so in fields involving complex or high dimensional data where model selection, or more generally predictor selection, is the main focus. Several examples of how the desiderata can be applied in practice to identify good predictors and assess their properties are given.

320 Statistical Methods for the Analysis of High Dimensional Data ■●

SSC, International Chinese Statistical Association, International Indian Statistical Association

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Model Selection for High-Dimensional Data with Applications in Feature Selection and Network Building

✦Xin Gao, York University, Department of Mathematics and Statistics, 4700 Keele Street, Toronto, ON M3J1P3 canada, *xingao@mathstat.yorku.ca*; Peter Song, University of Michigan; Yuehua Wu, York University

Key Words: composite likelihood, variable selection, Gaussian graphical model, BIC, model selection, consistency

For high-dimensional data set with complicated dependency structures, the full likelihood approach often leads to intractable computational complexity. This imposes difficulty on model selection as most of the traditionally used information criteria require the evaluation of the full likelihood. We propose a composite likelihood version of the Bayesian information criterion (BIC) and establish its consistency property for the selection of the true underlying marginal model. Under some mild regularity conditions, the proposed BIC is shown to be selection consistent, where the number of potential model parameters is allowed to increase to infinity at a certain rate of the sample size. In this talk, we will also discuss the result that using a modified Bayesian information criterion (BIC) to select the tuning parameter in penalized likelihood estimation of Gaussian graphical model can lead to consistent network model selection even when P increases with $N,$ as long as all the network edges are contained in a bounded subset.

Penalized Regression with Application to Genetic Association Studies

✦Wei Pan, University of Minnesota, Division of Biostatistics, Minneapolis, MN 55455 USA, *weip@biostat.umn.edu*

Key Words: Complex traits, Penalized regression, Rare variants, SNP, Statistical tests

Recent biotechnological breakthroughs have enabled large-scale genetic association studies to uncover common genetic variants predisposing to common and complex human diseases. However, recently published genome-wide association studies (GWASs) have confirmed the typically small to modest effect sizes of common genetic variants and limited statistical power of the standard single-marker analysis, thus it is critical to develop powerful statistical methods for multi-locus analyses to maximize the chance for discovery. We discuss novel and powerful penalized regression methods for analyses of common and rare variants for genetic association.

Statistical Methods For Integrative Genomics: Challenges And Opportunities

✦Joseph Beyene, McMaster University,

Due to rapid technological advances in recent years, various types of high-throughput genomic data with varying sizes, structures and complexities have become available. Among them are Single Nucleotide Polymorphisms (SNPs), Copy Number Variations (CNVs) and microarray gene expression measurements. Each of these distinct data types provides a different, partly independent and complementary view of the whole genome. However, understanding functions of genes and other aspects of the genome requires more information than provided by each of the data sets. I will present a conceptual integrative analysis framework and highlight novel statistical methods we have developed recently that can be used to integrate heterogeneous data types in order to answer different scientific questions. In particular, I will describe methods that can be used to answer questions involving class comparisons, quantifying associations between different sets of variables, and predicting clinical outcome. I will provide illustrative examples and discuss methodological issues and challenges.

321 Improving statistical literacy by international cooperation ●

International Association for Statistical Education, Section on Statistical Education, Scientific and Public Affairs Advisory Committee, Friends of Australasia, Statistics Without Borders

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Strategies for Stimulating Statistical Literacy and Understanding Quantitative Evidence in Higher Education in the UK

✦Gillian Lancaster, Lancaster University, Postgraduate Statistics Centre, Fylde College, Lancaster, International LA1 4YF UK, *g.lancaster@lancs.ac.uk*

Key Words: statistical literacy, quantitative methods, higher education, postgraduate

Statistical literacy plays an important part in our every day lives, helping us to make sense of news reports, magazine articles and health scares that can be over-sensationalised by the media. Yet there is currently a great need to encourage people to engage with, understand and question the quantitative information around us. Many courses on statistical methods are unpopular, as people cannot see the value of learning statistics and view it as a recipe collection of hard-to-grasp methods. The problem of disengagement is seen across the spectrum - in schools, in the workplace and in Higher Education, and this has stimulated a number of international initiatives including the Royal Statistical Society's 'getstats' campaign and the International Statistical Literacy Project (ISLP). Various frameworks have been proposed to stimulate learners' interest and take many guises. Problem-based learning for example can be both a data-driven process with little contextual motivation or a method of independent self-directed learning. This paper gives examples of some of the teaching strategies and networking initiatives undertaken by the Postgraduate Statistics Centre, Lancaster University.

Improving Statistical Literacy Through Graduate Programs in Ethiopia: The Case of North-South-South Collaborative Project in Jimma University

✦Yehenew Getachew, Jimma University, College of Agri & Vet Medicine, Jimma, P.O.Box 30 Ethiopia, *yehenew.getachew@ju.edu.et*; Luc Duchateau, Ghent University

Key Words: North-South-South, Statistical Literacy, Jimma University, Inter-University Program

Introduction: Jimma University (JU) is a public university mandated to run graduate and undergraduate programs. Recently, JU has ranked first among all public and private universities in Ethiopia. Challenges: In addressing the mandates, JU has faced chronic problems in the areas of Statistics, because of insufficient professional statisticians. Opportunities: As a source of professionals, JU critically searched for statisticians working in statistics offices, research centers and other universities. The existence of an Inter-University collaborative program with Belgian universities was the second opportunity. With aim of pulling together all these opportunities, JU proposed a NSS project in Statistics, with South partners, Jimma, Hawassa, AddisAbaba, Gonder and Mekele Universities from Ethiopia and Eduardo Mondlane University, from Mozambique, and North partners, Ghent, Hasselt, Leuven Universities from Belgium. A promising start: this collaborative initiative has resulted in running a joint graduate program, in various fields of applied statistics in JU and other sister universities; and also enabled us to offer various refresher statistical trainings (www.NSSbiostat.ugent.be).

Developing Statistical Literacy with Year 9 Students: A Collaborative Research Project

✦Sasha Sharma, The University of Waikato,

One of the most important goals for teaching statistics in schools today is to prepare students to deal with the statistical information that increasingly permeate their lives. More specifically, students should be able to read, understand and critically evaluate arguments and reports on a range of issues in a statistical manner arguing from and with data. However, the research on developing this statistical literacy is in its infancy. There is a need for empirical studies that investigate students' ability to evaluate social information. According to Bakker (2004), statistical ideas need to be developed slowly and systematically using carefully designed sequences of activities in appropriate learning environments. One way to develop these sequences of activities is through a research-and-development process called teaching experiment or design research (Cobb, 2002). Design research is cyclic with action and critical reflection taking place in turn. In this type of collaborative research, teachers and researchers are involved in the whole process and take part in posing questions, collecting data, drawing conclusions and writing reports. In our project, two cycles of teaching experiments were carried out in two Pasifika dominated classes with about 25 students. The following research questions guided the study: 1. How can we support students to develop statistical literacy within a data evaluation environment? 2. How can we develop a classroom culture where students learn to make and support statistical arguments based on data in response to a question of interest to them? 3. What learning activities and technology can be used in the classroom to develop students' statistical critical thinking skills? The teaching experiment had three phases: preparation, classroom teaching and interactions

with students, and debriefing and analysis of the teaching episodes. During the preparation phase the research team proposed a sequence of ideas, skills, knowledge and attitudes that they hoped students would develop as they participated in activities. The teaching phase took place in regular classrooms during normally scheduled mathematics lessons. The research team performed a retrospective analysis after each lesson to reflect on, revise conjectures and redirect the learning trajectory. The data set consisted of audio and video-recordings, copies of the students' written work, and field notes from the classroom sessions. Semi-structured interviews were also conducted with a selected number of students from each class while the design experiment is in progress. Each teacher-researcher kept a logbook of specific events that take place during the data collection period. Qualitative techniques are being used to analyse the data collected through interviews and observations made during the teaching sessions. The researchers intend to quantify students' responses to the survey data against a proposed statistical literacy framework based on the work of Watson and Callingham (2003) so as to triangulate the data. Quantitative analysis will measure students' movement across the five constructs of the proposed framework. Preliminary results show that literacy skills are critical for statistical literacy. This presents various demands on students' literacy skills. Classroom discourse is also important for developing statistical literacy in the classroom. Good and real data sets are needed to engage students' motivation. To discourage students from becoming too sceptical about statistics, it is important to provide examples where statistics are used correctly, Statistical literacy develops slowly and takes time. A careful sequence of activities can help develop an understanding of concepts such as sample and inference. This study has potential consequences in how the teaching of statistical literacy might be altered for greater effectiveness. The findings will contribute to the refinement of conceptual models developed in earlier research and assist teachers by providing a developmentally based hierarchy for teaching and assessing students. References Bakker, A. (2004). Reasoning about shape as a pattern in variability. Statistics Education Research Journal, 3(2), (pp. 64-83). Cobb, P. (2000). Conducting teaching experiments in collaboration with teachers. In A. Kelly & R. Lesh (Eds.), Handbook of research design in mathematics and science (pp. 307-333). Mahwah, NJ: Lawrence Erlbaum. Ministry of Education. (2007). Literacy Progressions: Meeting the Reading and Writing Demands of the Curriculum. Wellington: Learning Media Watson, J. M. and Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. Statistics Education Research Journal, 2(2), (pp. 3-46).

322 Computational Methods for Space-Time Correlated Data ■●

Section on Statistical Computing, Section for Statistical Programmers and Analysts, Section on Statistics and the Environment

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Using Approximate Dynamics to Inform Spatio-Temporal Ocean Ecosystem Models

✦Christopher K. Wikle, University of Missouri, Department of Statistics, 146 Middlebush Hall, Columbia, MO 65211-6100, *WikleC@Missouri.edu*

Key Words: dynamic, spatio-temporal, ocean, ecosystem, nonlinear, hierarchical

A wide variety of physical-biological models of varying complexity have been developed for components of the lower trophic ecosystem and its spatio-temporal interaction with the physical ocean. Such models are highly nonlinear and include parameterizations that are historically poorly informed by limited and relatively uncertain observations. This talk describes how approximate dynamical representations can facilitate parameter and state estimation for such processes in a Bayesian hierarchical framework.

Computationally Feasible Hierarchical Modeling Strategies for Large Spatial Data Sets

✦Sudipto Banerjee, University of Minnesota, 420 Delaware Street SE. MMC 303, Division of Biostatistics., Minneapolis, MN 55455 USA, *baner009@umn.edu*

Key Words: Bayesian modeling, Low-rank Gaussian processes, Hierarchical modeling, Markov chain Monte Carlo, Spatial data, Spatial super-populations

Large point referenced datasets are common in the environmental and natural sciences. The computational burden in fitting large spatial datasets undermines estimation of Bayesian models. We explore several improvements low-rank and other scalable spatial process models including reduction of biases and process-based modeling of ``centers'' or ``knots'' that determine optimal subspaces for data projection. We also consider alternate strategies for handling massive spatial datasets. One approach concerns developing process-based super-population models and developing Bayesian finite-population sampling techniques for spatial data. We also explore model-based simultaneous dimension-reduction in space, time and the number of variables. Flexible and rich hierarchical modeling applications in forestry are demonstrated.

Introducing Covariates in the Covariance Structure of Spatial and Spatio-Temporal Processes

✦Alexandra M. Schmidt, Federal University of Rio de Janeiro, Caixa Postal 68530, Rio de Janeiro, International 21.945-970 Brazil, *alex@im.ufrj.br*; Peter Guttorp, University of Washington; Joaquim Henriques Neto, Federal University of Rio de Janeiro; Anthony O'Hagan, University of Sheffield

Key Words: Anisotropy, Convolution, Manifold, Projection

In the analysis of most spatio-temporal processes underlying environmental studies there is little reason to expect spatial covariance structures to be stationary over the spatial scales of interest. This is because there may be local influences in the correlation structure of the spatial random process. Many alternatives to the usual stationary models have been proposed in the last decade, most of them based on highly stochastic systems. We discuss models for spatial covariance structures which relax the assumption of stationarity while keeping relative model simplicity. This is done by accounting for covariate information in the covariance structure of the spatial process. In particular, we discuss the inclusion of covariate information in the latent space approach of Sampson and Guttorp (1992) and Schmidt and O'Hagan (2003), and the convolution approach of Higdon (1998). We also developed applets

to visualize better the proposed nonstationary covariance structures. This is joint work with Peter Guttorp, Joaquim H. Viana Neto, and Tony O'Hagan.

323 Sequential Environmental Sampling ■

Section on Statistics and the Environment, ENAR, Section on Health Policy Statistics

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

The Use of Sequential Methods to Assess Trend in Florida Black Bear Populations

✦Linda J Young, University of Florida, Dept. of Statistics; IFAS, P.O. Box 110339; University of Florida, Gainesville, FL 32606, *LJYoung@ufl.edu*; Erin Leone, Florida Fish and Wildlife Conservation Commission; Brian Scheick, Florida Fish and Wildlife Conservation Commission

Key Words: SPRT, Sequential, 2-SPRT, spatial correlation

The Florida black bear subspecies (Ursus americanus floridanus) is listed as a threatened species by the state of Florida. The black bear population is monitored within the state and any trend in population size is of great interest. Because black bears are shy and secretive, direct observation is not possible, and hair traps are used to determine the presence or absence of a bear at a sampling location. Building on the work of Schipper and Meelis (1997 JABES; 2003 JABES), sequential tests for a trend in the black bear population size are compared and conducted. The impact of spatial and serial correlation is discussed.

Sequential Change-Point Detection in the Spread of Infectious Diseases

✦Michael Baron, University of Texas at Dallas, Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080, *mbaron@utdallas.edu*

Key Words: CUSUM, delay, detection, epidemic trend, false alarm, stopping time

Several outbreaks of rare infectious diseases and seasonal epidemics of influenza occurred in recent years. A developing area of statistical research is detection of changes in patterns of spread of infectious diseases. Timely detection of such changes can predict the beginning and magnitude of an impendent epidemic and provide important information for environmental health. A desired algorithm must attain sufficient sensitivity to detect changes early and fast, controlling at the same time the probability of false alarms. Epidemic models deal with stochastic processes that are marked by the presence of nuisance parameters, time-dependence, nonstationarity, and existence of rather complex prior information. Standard change-point detection methods are no longer optimal for such processes. Under these conditions, we formulate the change-point detection problem and propose stopping rules for its solution. They can be chosen in an optimal way, satisfying constraints on the average detection delay, the rate of false alarms, and probability of non-detection. Sequential tools are applied to the recent data published by the Centers for Disease Control and Prevention.

On a Class of Nonparametric Random Semi-Sequential Tests for Two-Sample Location Problem

✦AMITAVA MUKHERJEE, AALTO UNIVERSITY, SCHOOL OF SCIENCE AND TECHNOLOGY, Department of Mathematics and System Analysis, Otakaari-1M, PO BOX-11100, Room-335, Aalto, 00076 Finland, *amitmukh2@yahoo.co.in*

Key Words: Nonparametric Tests, Sequential Sampling, Arsenic Contamination, Monte-Carlo, Inverse Sampling, Semi-Sequential

In the present paper, we introduce a class of nonparametric two-sample tests based on a new semi-sequential sampling scheme. An existing partially sequential or semi-sequential procedure based on inverse sampling scheme, pioneered by Wolfe (1977) and Orban and Wolfe (1980), is modified in the light of random sequential sampling techniques, proposed by Mukhopadhyay and De-Silva (2008). Our proposed modification is motivated from a practical situation arise in geological field experiment. We discuss in detail the statistical methodologies and some asymptotic results. Numerical results based on Monte-Carlo are provided to justify asymptotic theory. Some power performances against fixed alternative are studied. We also provide an illustrative example with Arsenic contamination data.

324 Medallion Lecture: Recent developments in Bayesian methods for discovering regression structures: applications in the health sciences. ■●

IMS, ENAR, International Chinese Statistical Association, Section on Bayesian Statistical Science, WNAR

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Recent Developments in Bayesian Methods for Discovering Regression Structures: Applications in the Health Sciences

✦Sylvia Richardson, Department of Epidemiology and Biostatistics, Imperial College London, UK, , *sylvia.richardson@imperial.ac.uk*

In this talk, I shall review new developments and challenges related to the discovery of latent regression structures via Bayesian hierarchical models in a range of epidemiological and genetic problems. In the first part, the presentation will focus on outlining the construction and interpretation of flexible clustering structures within a regression framework aimed at capturing the effect of complex combination of predictors on health outcomes. In the second part, models and algorithms developed to discover sparse regression structures that may link several high dimensional data sets will be outlined. The talk will be illustrated throughout by case studies from epidemiology and integrative genomics.

325 Modern Data Analysis with Order Restrictions: A Tribute to Tim Robertson ■●

Council of Chapters, International Indian Statistical Association
Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Testing for or Against a Union-of-Cones Hypothesis with Applications to Genomic Data Analysis

✦Dan Nettleton, Iowa State University, , *dnett@iastate.edu*

Key Words: order restricted inference, multiple testing, microarray, gene expression

In some genomics applications, it is natural for either a null hypothesis or an alternative hypothesis to constrain a parameter vector to a union of cones. Although there is a long history of results from order restricted inference that apply when a parameter vector is constrained to a cone, relatively little work has been done for the union-of-cones case. We will discuss unique features of estimation and hypothesis testing when either the null hypothesis or the alternative hypothesis constrains a parameter vector to a union of cones. Two example applications from genomics will be presented. The first application involves identifying differentially expressed gene sets by searching for monotonic changes in multivariate gene expression distributions. This is relevant for experiments were the multivariate expression distribution of the set of genes is measured for each level of a naturally ordered covariate like time or dose of a drug. A second application involves identifying genes that show evidence of heterosis (hybrid vigor) for gene expression traits when analyzing data from parental inbred lines and their hybrid offspring.

Order-Restricted Inference: Computational Algorithms

✦Edward J Wegman, George Mason University, MS 6A2, 4400 University Drive, Fairfax, VA 22030 USA, *ewegman@gmail.com*

Key Words: order-restricted inference, isotonic regression, computational methods

Order-restricted inference gained much currency in the 1960s in an era where mainframe computing was the norm. In the last forty-five years, computation has made enormous strides. In the paper I will review approaches to computational algorithms for order-restricted inference including Bayesian methods, R-based algorithms, and suggest some hybrid methods. This presentation is dedicated to the memory of Professor Tim Robertson.

Comparison of Two Nonparamatric Regression Curves: Test of Superiority and Noninferiority

✦Mervyn Joseph Silvapulle, Monash University, Dept Econometrics and Bus Stats, P. O Box 197, Caulfield East, International 3145 Australia, *Mervyn.Silvapulle@monash.edu*

Key Words: Constrained inference, Noninferiority, Order restricted inference, Superiority

Tests are developed for detecting differences between two univariate nonparametric regression curves. The objective of the new method is to establish that a treatment is noninferior to another for the whole population and also that it is superior at least for a part of the population, when the treatment effect is represented by a nonparametric regression curve. The inference problem is formulated as test against the alternative hypothesis which says that (a) the regression curve for population 1 does not fall below that for population 2 by more than a specified small amount at any value of the covariate, and (b) the former exceeds the latter, at some values of the covariate, by more than a specified amount. The test statistic is easy to compute and also apply using a table of asymptotic critical values. Because this test is conservative, a less conservative bootstrap test is proposed and is shown to be asymptotically valid. In a simulation study, we observed that the type I error rates for these tests were close to the nominal level, and the bootstrap test exhibited higher estimated power, as expected.

326 Understanding Intervention Mechanisms: Causal Mediation Analysis and Principal Stratification ■●

Social Statistics Section, Section on Health Policy Statistics
Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Conventional and Principal Stratification Perspectives on Mediation

✦Booil Jo, Stanford University, Dept. of Psychiatry and Behavioral Sciences, 401 Quarry Rd., MC5795, Stanford, CA 94305-5795 USA, *booil@stanford.edu*; Elizabeth A Stuart, Johns Hopkins Bloomberg School of Public Health

Key Words: Baron and Kenny approach, McArthur approach, Principal stratification, Causal inference, Mediation, Conditional ignorability

We will first review how mediation is understood in conventional approaches such as the Baron/Kenny and the McArthur approach. This is an important step given that these approaches are widely used in many areas of social, behavioral, and medical research. Then, we will clarify how their underlying assumptions are interpreted in contemporary causal mediation approaches such as principal stratification. Principal stratification refers to cross-classification of individuals based on potential values of posttreatment variables under all compared treatment conditions. Since the resulting strata are unaffected by treatment, treatment effects conditioning on principal strata can be interpreted as causal effects. Finally, we will focus on the conditional ignorability (i.e., conditional on pretreatment covariates, individuals with different mediator status are comparable), which is the central assumption necessary for causal interpretation in conventional mediation analyses, and is often strongly criticized. We will examine implications of this assumption from the principal stratification perspective and compare the utility of the assumption in conventional and causal mediation modeling.

Experimental Designs for Identifying Causal Mechanisms

✦Kosuke Imai, Princeton University, Department of Politics, Princeton University, Princeton, NJ 08540, *kimai@princeton.edu*; Dustin Tingley, Harvard University; Teppei Yamamoto, Princeton University

Key Words: causal inference, causal effects, indirect effects, direct effects, identification, causal mediation

Experimentation is a powerful methodology that enables scientists to empirically establish causal claims. However, one important criticism is that experiments merely provide a black-box view of causality and fail to identify causal mechanisms. Critics argue that although experiments can identify average causal effects, they cannot explain how such effects come about. If true, this represents a serious limitation of experimentation, especially for social and medical science research whose primary goal is to identify causal mechanisms. In this paper, we consider several different experimental designs and compare their identification power. Some of these designs require the direct manipulation of mechanisms, while others can be used even when only imperfect manipulation is possible. We use recent social science experiments to illustrate the key ideas that underlie each design.

Augmented Designs to Assess Principal Strata Effects

✦Fabrizia Mealli, Department of Statistics - University of Florence, Florence, 50139 Italy, *mealli@ds.unifi.it*; Alessandra Mattei, Department of Statistics - University of Florence

Key Words: Augmented Designs, Bounds, Causal Inference, Principal Stratification, Direct and Indirect Effects

Many research questions involving causal inference are often concerned with understanding the causal pathways by which a treatment affects an outcome. We tackle the problem of disentangling direct and indirect effects by investigating new augmented experimental designs, where the treatment is randomized, and the mediating variable is not forced, but only randomly encouraged. There are two key features of our framework: we adopt a principal stratification approach, and we mainly focus on principal strata effects, avoiding to involve a priori counterfactual outcomes. Using non parametric identification strategies, we provide a set of assumptions, which allow us to partially identify the causal estimands of interest: the Principal Strata Direct Effects. Some examples are shown to illustrate our design and causal estimands of interest. Large sample bounds for the Principal Strata average Direct Effects are provided, and a simple hypothetical example is used to show how our augmented design can be implemented and how the bounds can be calculated. Finally our augmented design is compared with and contrasted to a standard randomized design.

Mixture Modeling of Treatment Effects with Multiple Compliance Classes and Missing Data

✦Michael Sobel, Columbia University, 10027, *mes105@columbia.edu*; Bengt Muthen, University of California,Los Angeles

Key Words: causal inference, complier average causal effect

Randomized experiments are the gold standard for making causal inferences. Researchers design treatments to affect mediators lying on one or more presumed pathways to the outcome. Investigators typically want to know the effect of offering the treatment and also the effect of the treatment itself. To address the latter question, recent attention has focused on the effect among subjects who will comply with their treatment assignment. In many cases, there is little reason to believe that the mediators targeted by the treatment will produce effects for all complier subjects. Therefore, we estimate the proportion of compliers unaffected by treatment as well as the proportion affected and the effect. Missing data further complicate estimation and we consider various missing data assumptions, including the assumption that the missing data are missing at random and the assumption of latent ignorability.

327 Sparse Regression and High-dimensional Statistical Analysis ●

IMS, International Chinese Statistical Association, Section on Statistical Computing

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Exponential Screening and Optimal Rates of Sparse Estimation

✦Philippe Rigollet, Princeton University, Operations Research and Financial Engineering, Sherrerd Hall, Princeton, NJ 08544, *rigollet@princeton.edu*; Alexandre Tsybakov, Laboratoire de Statistique, CREST

Key Words: high-dimensional regression, aggregation, adaptation, sparsity, sparsity oracle inequalities, minimax rates

We consider a general, non necessarily linear, regression problem with Gaussian noise and study an aggregation problem that consists in finding a linear combination of approximating functions, which is at the same time sparse and has small mean squared error (MSE). We introduce a new estimation procedure, called Exponential Screening (ES) that shows remarkable adaptation properties: it adapts to the linear combination that optimally balances MSE and sparsity, whether the latter is measured in terms of the number of non-zero entries in the combination or in terms of the global weight of the combination. The power of this adaptation result is illustrated by showing that ES solves optimally and simultaneously all the problems of aggregation in Gaussian regression considered previously. Tight minimax lower bounds establish optimal rates of sparse estimation and that the ES procedure is optimal. Finally, a numerical implementation of ES that results in a stochastic greedy algorithm is discussed and compared to state-of-the-art procedures for sparse estimation.

Global Testing Under Sparse Alternatives: Anova, Multiple Comparisons, and the Higher Criticism

Ery Arias-Castro, Department of Mathematics, UCSD; ✦Yaniv Plan, California Institute of Technology, Pasadena, CA 91125, *plan@caltech.edu*; Emmanuel J Candes, Department of Statistics, Stanford University

Key Words: Detecting a sparse signal, analysis of variance, higher criticism, minimax detection, compressive sensing, incoherence

We study the problem of testing for the significance of a subset of regression coefficients in a linear model under the assumption that the coefficient vector is sparse, a common situation in modern high-dimensional settings. Assume there are p variables and let S be the number of nonzero coefficients. Under moderate sparsity levels, when we may have S > p^(1/2), we show that the analysis of variance F-test is essentially optimal. This is no longer the case under the sparsity constraint S < p^(1/2). In such settings, a multiple comparison procedure is often preferred and we establish its optimality under the stronger assumption S < p^(1/4). However, these two very popular methods are suboptimal, and sometimes powerless, when p^(1/4) < S < p^(1/2). We suggest a method based on the Higher Criticism that is essentially optimal in the whole range S < p^(1/2). We establish these results under a variety of designs, including the classical (balanced) multi-way designs and more modern `p > n' designs arising in genetics and signal processing.

328 Teaching Regression through Statistics ■

Section on Statistical Education

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Teaching Statistics Through Regression

✦Felicity Enders, Mayo Clinic, 200 First St. SW, Rochester, MN 55905, *enders.felicity@mayo.org*; ✦Shonda Kuiper, Dept. of Math and Statistics, Grinnell College, Grinnell, IA 50112-1616, *kuipers@math.grinnell.edu*; ✦Daniel T Kaplan, Macalester College, 1600 Grand Ave., Saint Paul, MN 55105 USA, *kaplan@macalester.edu*; ✦Laura Sather Zielgler, Dept. of Statistics, Saint Cloud State University, Engineering and Computing Center 141, Saint Cloud, MN 56301-4498, *lsather@stcloudstate.edu*

Key Words: regression, modeling, education, undergraduate

A standard introductory statistics course barely touches techniques that are the bread and butter of practicing statisticians. One strategy that can help to align better practice with teaching is to emphasize modeling and regression. The panelists will describe their experiences doing this at different levels in the undergraduate curriculum, the challenges involved and the benefits that accrue, and ways to assess statistical learning in a regression-based approach.

329 Rating Competitors in Games and Sports in the 21st Century ■

Section on Statistics in Sports

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

The Abcs Of Xqjkz: A New Scrabble Rating System Based On A Statistical Model For Tile-By-Tile Play

✦Kenneth Shirley, AT&T Labs Research, 180 Park Ave, Building 103, Florham Park, NJ 07932, *kshirley@research.att.com*

Key Words: Scrabble, Rating system, Hierarchical Bayes

We develop a statistical model for Scrabble in which we model the number of points scored on each turn as a function of the individual tiles in a player's rack. The result is a detailed model that describes a player's Scrabble skill in terms of dozens of player-specific variables related to interpretable Scrabble skills such as how often a player plays each tile, how many points he earns per tile, and how much he augments his score by incorporating tiles already on the board into his play. Our data comes from a public database of about 600 games of Scrabble played at the expert level. We find that most of the variation in points scored is explained by the frequency with which tiles are played and the frequency with which players get a "bingo" (playing all 7 tiles in the rack on a single turn). The player-specific model parameters can be used as the basis of a much more detailed player rating system than is currently in use. This work also sheds light on the degree to which the outcomes of Scrabble games depend on the randomness inherent in drawing tiles, as opposed to player skill. The largest component of the model consists of a hierarchical Bayesian logistic regression model.

Ratings Par Excellence: A New Handicapping System for Golf

✦Tim Swartz, simon fraser university, dept of stats/actsci, 8888 university drive, burnaby, BC v5a1s6 canada, *tim@stat.sfu.ca*

Key Words: data analysis, golf, handicapping, normal distribution, order statistics

The handicapping system of the Royal Canadian Golf Association (RCGA) is very similar to the handicapping system of the United States Golf Association (USGA). Although these handicapping systems are complex and have been carefully studied, the systems do not take statistical theory into account. In 2000, the Handicap Research Committee of the RCGA was formed and challenged with the task of developing a new handicapping system. This talk outlines the proposed system. The proposed system continues to make use of the existing course ratings and slope ratings, but uses statistical theory to drive the methodology. In this talk, we demonstrate that the proposed system has several advantages over the current system, including fairness and improved interpretability. The proposed system is supported by both theory and data analyses.

Paired Comparison Models with Tie Probabilities and Order Effects as a Function of Strength

✦Mark E. Glickman, Boston University School of Public Health, EN Rogers Memorial Hospital (152), Building 70, 200 Springs Road, Bedford, MA 01730 USA, *mg@bu.edu*

Key Words: tournament, chess, DIC, Bayesian

Paired comparison models, such as the Bradley-Terry model and its variants, are commonly used to measure competitor strength in games and sports. Extensions have been proposed to account for order ef-

fects (e.g., home-field advantage) as well as the possibility of a tie as a separate outcome, but such models are rarely adopted in practice due to poor fit with actual data. We propose a novel paired comparison model that accounts not only for ties and order effects, but recognizes two phenomena that are not addressed with commonly used models. First, the probability of a tie may be greater for stronger pairs of competitors. Second, order effects may be more pronounced for stronger competitors. This model is motivated in the context of tournament chess game outcomes. The models are demonstrated on the 2006 US Chess Open, a large tournament with players of wide-ranging strengths, and to the Vienna 1898 chess tournament, a double-round robin tournament consisting of 20 of the world's top players.

Statistics in the Ballroom: Rankings, Voting, Fairness, Prediction, and Quantifying the Palin Effect on 'Dancing with the Stars'

✦Jason A. Gershman, Nova Southeastern University, 3301 College Avenue, MCT: Mailman BLDG 2nd Floor, Fort Lauderdale, FL 33314, *jgershma@nsu.nova.edu*

Key Words: Fairness, Sports, Voting, Rankings, Ballroom, Dancing

The popular ballroom dancing competition television show Dancing With The Stars (DWTS) utilizes a ranking system with some intriguing statistical aspects. Some shortcomings of the ranking method were well publicized following Bristol Palin's unlikely position in the finale of Season 11 of the show. This paper examines the voting strategy which led to this unlikely result. Examining fairness and technique, I will contrast the DWTS ranking system with other ranking and voting systems such as those used to elect presidents, decide which college football teams compete for the national championship, and determining the next American Idol. I also examine optimal voting strategies and game theory within the context of real situations from this ballroom dancing competition. Finally, I also demonstrate how an extremely biased sample can be used to make accurate unbiased inference using freely available polling data from fans utilizing the most popular voting method for electing the celebrity ballroom dancing champion.

Ratings Central: Accurate, Automated, Bayesian Table Tennis Ratings for Clubs, Leagues, Tournaments, and Organizations

✦David J Marcus, Ratings Central, 25 Beacon St Apt 16, Somerville, MA 02143-4336, *davidmarcus@alum.mit.edu*

Key Words: sports, ratings, rankings, Bayesian, table tennis, Ping-Pong

Ratings Central (www.ratingscentral.com) has been providing ratings for table tennis players in the U.S. and several other countries since 2004. Players range from five-year-olds to world champions. Currently, there are over 31,000 players and 660,000 matches. The system uses a Bayesian model with event directors providing priors for unrated players, specifying them either for a group (e.g., league division, tournament, age, gender) or individually. To make the model computationally tractable, the graph of matches in an event (nodes for players, edges for matches) is modified for each player while still retaining most of the information on how a player's opponents did in the event. The system occasionally has trouble dealing with special subpopulations or players whose playing strength jumps. Although players and event directors sometimes have misconceptions about the rating system, in general they are very satisfied.

330 Statistical Challenges and Advances in Epidemiological Studies with Risk Set Sampling ■●

Biometrics Section, ENAR, Section on Risk Analysis, Section on Statistics in Epidemiology, WNAR

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Joint Analysis Of Accelerated Failure Time And Longitudinal Data Under A Case-Cohort Design

✦Lan Kong, University of Pittsburgh, 15261, *lkong@pitt.edu*; Xinxin Dong, University of Pittsburgh

Key Words: Joint model, biomarker, mixed effects model, case-cohort

A panel of biological markers is often measured over time to better understand the mechanism of a disease and aid in the development of effective treatments. In large cohort studies, it is prohibitive to measure multiple candidate markers over time for each individual. Case-cohort design provides a cost effective solution when the covariate of interest is expensive to measure and the event rate is low. Under the case-cohort design, biomarkers are measured for a subcohort that is randomly selected from the entire cohort and any additional cases outside the subcohort. Joint modeling of longitudinal marker and clinical outcome is an appealing technique to reveal the relationship between biomarker trajectory and the evolution of the disease. We propose a joint analysis of accelerated failure time (AFT) and longitudinal data from case-cohort studies. We use the shared latent parameters to link the longitudinal model and AFT model, and obtain the maximum likelihood estimators by Gaussian quadrature method. We evaluate the performance of our case-cohort estimator and its relative efficiency to full cohort estimator through simulation studies and give a numerical example for illustration.

Biomarker Study Under Nested Case-Control Design

✦Yingye Zheng, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. NE, M2-B500, Seattle, WA 98109, *yzheng@fhcrc.org*; Tianxi Cai, Harvard University

Nested case-control design provides a cost-effective cohort sampling strategy in biomarker research. We propose estimating procedures to evaluate the prognostic accuracy of a novel biomarker for predicting future events with censored failure time outcome under a matched nested case-control design. In the presence of missing information on the biomarker measurements, we develop consistent estimators for time-dependent classification and predictive accuracy measures based on the idea of inverse probability weighting.

Group Lasso In Nested Case-Control Studies

✦Mengling Liu, New York University, , *mengling.liu@nyu.edu*

Key Words: Nested case-control study, Risk-set sampling, Spline regression, Variable selection

Prominent for its merit of being cost effective in studying the association between disease and its risk factors, the nested case-control (NCC) design has been commonly used in large cancer epidemiology studies. When a large number of genetic, environmental, and clinical covariates are available, the dimension of predictors under consideration can easily grow enormously. Moreover, effects of risk factors on the disease may be nonlinear. In this paper, we develop an effective variable selection procedure to building models that provide better risk assessment and model interpretation with NCC data using the group lasso technique. Extensive simulations are conducted to evaluate the finite sample performance of our proposed approach. We further demonstrate the proposed method with data from an NCC study in the New York University Women's Health Study.

Efficiency Of The Profile Likelihood Estimator For Case-Control Studies Under General Misspecification

✦Jennifer L Wilcock, University of Auckland, Department of Statistics, Private Bag 92019, Auckland, International 1142 New Zealand, *j.wilcock@auckland.ac.nz*; Alan J Lee, University of Auckland

Key Words: case-control, semiparametric efficiency, asymptotic variance

Scott and Wild (1997 and 2001) introduced a profile likelihood approach to estimation for generalised case-control studies and the method has recently been extended to accommodate multi-phase case-control designs (2010). In this talk we describe a simple and general approach which can be used to demonstrate the semiparametric efficiency of the Scott-Wild estimator in a variety of response-selective schemes whilst also allowing for a general misspecification of the model. Our approach is based on projections and an adaptation of the method of Newey (1994). Previous work is extended in two ways. First, we treat the case of efficiency in a wide-sense where the estimation is carried out under possible general misspecification rather than under the assumption that the model is true. Second, our approach allows a variety of contexts to be handled by a single argument rather than each context requiring extensive specialised calculations as in Lee and Hirose (2008).

Efficiency Loss When Using Frequency Matching And Balance As Design Strategies In Case-Control And Two-Phase Studies

✦Alan J Lee, University of Auckland, Department of Statistics, Private Bag 92019, Auckland, International 1142 New Zealand, *aj.lee@auckland.ac.nz*; Jennifer L Wilcock, University of Auckland

Key Words: Balanced design, Frequency matching, Two-stage design, Optimal design, case-control

Frequency matching is a common design strategy aimed at improving the efficiency of case-control studies, while the use of balanced designs is a similar strategy in the case of two-phase designs. In this paper, we examine the efficiency loss (compared to the theoretical optimal design) when pursuing these strategies. We assume that estimation is done using the semi-parametric method of Scott and Wild (Scott and Wild, Biometrika,1997), and develop an asymptotic formula for the variance of the estimator in both these designs. This is used to find optimal designs and hence compute the efficiency loss. Some extensions to related situations such as multistage designs are considered.

331 The Supplemental Poverty Measure: New Research Findings ■●

Section on Government Statistics, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Supplemental Poverty Measure Thresholds: Imputing Noncash Benefits To The Consumer Expenditure Survey Using The Current Population Survey

✦Thesia Garner, Bureau of Labor Statistics, Postal Square Bldg., #3105, 2 Mass. Ave., NE, Washington, DC 20212, *garner.thesia@bls.gov*; Charles Hokayem, U.S. Census Bureau

Key Words: poverty measurement, thresholds, in-kind benefits, imputation, Consumer Expenditure Survey, Current Population Survey

In March 2010 an Interagency Technical Working Group (ITWG) released guidelines on thresholds and resources for a Supplemental Poverty Measure (SPM). The ITWG recommended that thresholds include in-kind benefits that are accounted for in resources; however, only limited in-kind benefit information is available in the data set upon which the thresholds are based, the Consumer Expenditure Survey (CE). For example, the CE collects information on food expenditures that implicitly include the cash value of benefits from the Supplemental Nutrition Assistance Program but no information on other food programs. In earlier work, Garner (2011) imputed in-kind rates and benefits for school lunches and WIC to the CE using eligibility guidelines. However, eligibility does not equal actual or reported participation rates. To better reflect reported rates of participation, data from the Current Population Survey (CPS), the basis of the SPM resource measure, are used to model imputations to the CE for participation in school lunches and WIC. Methods used to assign the in-kind benefit levels to the CPS are used to assign values to the CE. Rates from this and the Garner study are compared.

A Unit Of Analysis For Poverty Measurement

✦Ashley Provencher, U.S. Census Bureau, , *ashley.provencher@census.gov*

In 2009 the Office of Management and Budget's Chief Statistician formed an Interagency Technical Working Group (ITWG) on Developing a Supplemental Poverty Measure. In March 2010 the ITWG issued a series of suggestions on how to develop a new measure drawing on the recommendations of the 1995 report of National Academy of Sciences (NAS) Panel on Poverty and Family Assistance and the extensive research on poverty measurement conducted over the past 15 years. One suggestion of the ITWG was that the family unit should be broadened to include all related individuals who live at the same address, any co-resident unrelated children who are cared for by the family (such as foster children), plus cohabitors and their children. This paper will examine how the change in unit of analysis from the family

definition used in the official poverty measure (a group of two or more people residing together related by birth, marriage, or adoption) to the broader definition impacts the incidence of poverty and composition of family units. The analysis will use data from the 2010 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC).

Medical Out-Of-Pocket Spending Among The Uninsured: Differential Spending & The Supplemental Poverty Measure

✦Kyle J Caswell, Bureau of the Census, DC , *kyle.j.caswell@census.gov*; Kathleen S Short, US Census Bureau

Key Words: Uninsured, Predictive Mean Matching, Medical Out-of-Pocket Spending, Poverty

This paper refines the treatment of Medical Out-of-Pocket (MOOP) spending among the uninsured in measuring poverty, and investigates its net effect on the Supplemental Poverty Measure (SPM). It extends previous research that accounts for low MOOP spending among the uninsured using the Current Population Survey 2010 Annual Social and Economic Supplement (CPS ASEC) and predictive mean matching methods. This work, like the former, is a direct response to recommendations of the Interagency Technical Working Group (ITWG) on developing a SPM in that it investigates the pros and cons of accounting for low MOOP spending on behalf of the uninsured. This work extends the former by making less restrictive assumptions on insurance coverage type in estimating counterfactual distributions of non-premium and premium MOOP spending for the uninsured. Additionally, premium expenditures are simulated for the uninsured using the 2014 provisions of the Patient Protection and Affordable Care Act. Finally, this work incorporates these refined counterfactual distributions of MOOP spending with the full range of ITWG recommendations to implement the SPM.

Research On Commuting Expenditures And Geographic Adjustments In The Supplemental Poverty Measure

✦Melanie A. Rapino, US Census Bureau, 4600 Silver Hill Road, 7H462C, Washington, DC 20233, *melanie.rapino@census.gov*; Brian McKenzie, US Census Bureau; Matthew Marlay, US Census Bureau

Key Words: supplemental poverty measure, poverty, commuting

The current SPM adjusts poverty thresholds for geographic difference based solely on differences in housing costs, in large measure because of the current limitations in data related to other costs. In Experimental Poverty Measures: 1999 (U.S. Census Bureau 2001), the 1995 National Academy of Sciences (NAS) Panel on Poverty and Family Assistance proposed subtracting a flat amount from a family's resources for 'other' work-related expenses, with an annual inflation adjustment. In future poverty measures, the JTWMSB recommends that commuting expenses be delineated by geography for a more accurate calculation of the SPM thresholds. This research will examine the current state of the SPM and provide steps for improvement of the measure. First, several U.S. government sponsored surveys will be examined for their potential assistance to geographically adjust the SPM and take into account commuting expenditures in the index. Surveys to be examined include: the Survey of Income and Program Participation (SIPP), Current Population Survey (CPS), National Household Travel Survey (NHTS), and the American Community Survey (ACS). Next, further research steps aimed at developing a more refi

332 Evaluation of Disclosure Limitation Methods for Medical Records and Claims Data as a Means to Increasing Access for Researchers

Section on Health Policy Statistics, Committee on Privacy and Confidentiality, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Evaluation And Limitation Of Disclosure Risk Of Health Data Using Random Substitution And Subsampling

✦Joshua Michael Borton, NORC at the University of Chicago, 55 East Monroe Street, 30th Floor, Chicago, IL 60603, *borton-joshua@norc.org*; Peter K Kwok, NORC at the University of Chicago; Avinash C Singh, NORC at the University of Chicago

Key Words: Statistical disclosure limitation, inside intrusion, de-identification, non-synthetic treatment, disclosure risk measure, data utility

In order for medical data, such as electronic medical records (EMR), to be shared for research purposes the data provider must be certain that the risk of re-identification of such records is sufficiently low. Statistical disclosure limitation (SDL) is the process of reducing disclosure risk. In order for the data to be useful to researchers the chosen SDL method should have as little impact as possible on inferences made from the data. It is the constant tension between data confidentiality and data quality that is at the heart of all decisions regarding SDL. We present a method by which both data quality and data confidentiality are evaluated throughout the process. This information is used to determine the amount of random substitution and subsampling that should be used as part of our SDL process. Abridged EMR data was used for this exercise.

Harder Than You Think- A Case Study Of Re-Identification Risk Of Hipaa-Compliant Records

✦Peter K Kwok, NORC at the University of Chicago, 55 East Monroe Street, 30th floor, Chicago, IL 60603, *kwok-peter@norc.org*; Michael Davern, NORC at the University of Chicago; Elizabeth C Hair, NORC at the University of Chicago; Deborah Lafky, Office of the National Coordinator for Health Information Technology

Key Words: Statistical disclosure limitation, HIPAA, Safe Harbor, re-identification, outside intrusion

We have studied the admission records of Hispanics in one hospital system between 2004 and 2009. The data set was stripped of identifying information as required by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) safe harbor methodology. We simulated an intrusion scenario in which an intruder had access to a substantial amount of information available from a market research company. We used the market research data to try to identify specific

people from the hospital system's HIPAA de-identified data set, and sent possible matches to the hospital system for confirmation. Our experiment shows that this intrusion scenario involves many challenges. Even when the intruder is given strong assumptions about their knowledge the re-identification risk is only about 0.22%. We discuss the limits of our analysis and identify areas for future inquiry.

Preparing Medicare Claim Data For Stochastic Statistical Disclosure Limitation Treatment

✦Tzy-Chyi Yu, NORC at the University of Chicago, 4350 East-West Highway, 8th Floor, Bethesda, MD 20814, *yu-tzy-chyi@norc.org*; Elizabeth C Hair, NORC at the University of Chicago; Beth A Virnig, University of Minnesota School of Public Health

Key Words: Comparative effectiveness research, Medicare claims, statistical disclosure limitation, data utility, inside intrusion

Comparative effectiveness research (CER) analysts prefer data at the individual level. However, data sets containing information required by CER researchers often risk disclosing sensitive information about individuals in the data through indirect identifiers. Statistical disclosure limitation (SDL) methods are used to reduce this risk, while maintaining analytic utility. When preparing data for SDL it is often necessary to aggregate data from multiple records, within numerous tables to create a single record describing the disc losable entity in detail. In the case of Medicare claims this requires an approach that scans records from multiple claim tables (inpatient, outpatient, prescription drug, etc.) to create a beneficiary profile representative of what an intruder could use to disclose sensitive information.

Nuts And Bolts Of Preparing A System For Summarizing Medicare Beneficiaries

✦Al Crego, NORC at the University of Chicago, 55 East Monroe Street, 30th Floor, Chicago, IL 60603, *crego-allan@norc.org*; Tzy-Chyi Yu, NORC at the University of Chicago; Joshua Michael Borton, NORC at the University of Chicago; Peter K Kwok, NORC at the University of Chicago

Key Words: Medicare claims, beneficiary-level information, processing techniques

Some CER needs having been determined to be at the level of unique beneficiaries, the challenge was to implement a summary system that was accurate, reproducible, yielded files in forms optimal for treatment procedures, and simultaneously served potential research needs. This was accomplished for two random, five-percent samples of 2008 Medicare beneficiaries. Sufficient flexibility was also necessary to accommodate changes to perceived or future research goals. We began with two databases, each comprised of over three-thousand variables and over one-hundred million records across various files, and developed a processing stream that could be applied, with minimal changes, to any number of similar databases, resulting in tractable summary files. In doing so, we applied industry-standard summary techniques in multiple dimensions, dependent on claims and beneficiary-level information.

Creation Of Public Use Files: Lessons Learned From The Comparative Effectiveness Research Public Use Files Data Pilot Project

✦Erkan Erdem, IMPAQ International LLC, 10420 Little Patuxent Parkway, Suite 301, Columbia, MD 21044, *eerdem@impaqint.com*; Sergio Prada, IMPAQ International LLC

Key Words: Public use files, re-identification, de-identification, Medicare claims, comparative effectiveness research, data utility

In this paper we describe the lessons learned from the creation of Basic Stand Alone (BSA) Public Use Files (PUFs) for the Comparative Effectiveness Research Public Use Files Data Pilot Project (CER-PUF). CER-PUF is aimed at increasing access to CMS claims data sets through the creation of public use files that: do not require user fees and data use agreements, have been de-identified to assure the confidentiality of the beneficiaries and providers, and provide analytic utility to researchers. This paper describes the steps taken in the project to strike the right balance between data utility and privacy protection. We draw lessons learned from three tasks: (i) the creation of each PUF involving design of the sample data, analysis of variables, analysis of de-identication strategies (including deterministic and stochastic treatment), risk analysis, and documentation, (ii) environmental scan including stake-holder interviews, case-studies of de-identified individual level public use data, and literature review and legal analysis, and (iii) review of the needs of comparative effectiveness researchers and statistical de-identification methods that are acceptable to them.

333 Nonparametric Section Student Paper Competition ●

Section on Nonparametric Statistics

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Robust Estimation For Homoscedastic Regression In The Secondary Analysis Of Case-Control Data

✦Jiawei Wei, Texas A&M University, 3143 TAMU, Department of Statistics, College Station, TX 77843-3143, *wjw@stat.tamu.edu*; Raymond James Carroll, Texas A&M University; Ursula U. M®uller, Texas A&M University; Ingrid Van Keilegom, UniversitÈ catholique de Louvain; Nilanjan Chatterjee, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, USA

Key Words: Biased samples, Homoscedastic regression, Secondary data, Secondary phenotypes, Semiparametric inference, Two-stage samples

Primary analysis of case-control studies focuses on the relationship between disease (D) and a set of covariates of interest (Y,X). A secondary application of the case-control study, often invoked in modern genetic epidemiologic association studies, is to investigate the interrelationship between the covariates themselves. The task is complicated due to case-control sampling. Previous work has assumed a parametric distribution for Y given X and derived semiparametric efficient estimation and inference without any distributional assumptions about X. In this paper, we take up the issue of estimation of a regression function when Y given X follows a homoscedastic regression model, but otherwise the distri-

bution of Y is unspecified. The semiparametric efficient approaches can be used to construct semiparametric efficient estimates, but they suffer from a lack of robustness to the assumed model for Y given X. We take an entirely different and novel approach in the case that the disease is rare. We show how to estimate the regression parameters in the rare disease case even if the assumed model for Y given X is incorrect, and thus the estimates are model-robust.

Robust Estimation Of Generalized Additive Models

✦Raymond Ka Wai Wong, Department of Statistics, University of California, Davis, 1 Shields Avenue, Davis, CA 95616, *raymond.kww@gmail.com*; Fang Yao, University of Toronto; Thomas C. M. Lee, Department of Statistics, University of California, Davis

Key Words: Bounded score function, Generalized information criterion, Generalized linear model, Robust estimating equation, Robust quasi-likelihood, Smoothing parameter selection

This article studies M-type estimators for fitting generalized additive models in the presence of anomalous data. A new theoretical construct is developed to link the costly M-type calculations with least-squares type computations. Its asymptotic properties are studied and used to motivate a computational algorithm. The main idea is to decompose the overall M-type estimation problem into a sequence of well-studied conventional additive model fittings. The resulting algorithm is fast and stable, can be paired with different nonparametric smoothers, and can also be applied to cases with multiple covariates. As another contribution of this article, automatic methods for smoothing parameter selection are proposed. These methods are designed to be resistant to outliers. The empirical performance of the proposed methodology is illustrated via both simulation experiments and real data analysis.

Fixed And Random Effects Selection In Nonparametric Additive Mixed Models

✦Randy C.S. Lai, UC Davis, 4118 Mathematical Sciences Building, One Shields Avenue, Davis, CA 95616, *rcslai@ucdavis.edu*

Key Words: adaptive group lasso, additive mixed model, Bayesian information criterion, consistency, oracle property

This paper considers the problem of model selection in a nonparametric additive mixed modeling framework. The fixed effects are modeled nonparametrically using truncated series expansions with B-spline basis. Estimation and selection of such nonparametric fixed effects are simultaneously achieved by using the adaptive group lasso methodology, while the random effects are selected by a traditional backward selection mechanism. To facilitate the automatic selection of model dimension, computable expressions for the degrees of freedom for both the fixed and random effects components are derived, and the Bayesian Information criterion (BIC) is used to select the final model choice. Theoretically it is shown that this BIC model selection method possesses the so-called oracle property, while computationally a practical algorithm is developed for solving the optimization problem involved. Simulation results show that the proposed methodology is often capable of selecting the correct significant fixed and random effects components, especially when the sample size is not too small.

A General Framework for Sequential and Adaptive Methods in Survival Studies

✦Gongjun Xu, Columbia University, New York, NY 10027 US, *gongjun@stat.columbia.edu*; Xiaolong Luo, Celgene Corporation; Zhiliang Ying, Columbia University

Key Words: Sequential analysis, Adaptive designs, the Cox model, Marked point process

Adaptive treatment allocation schemes based on interim responses have generated a great deal of recent interest in clinical trials and other follow-up studies. An important application of such schemes is in survival studies, where the response variable of interest is time to the occurrence of certain event. Due to possible dependency structure inherited in the enrollment and allocation schemes, existing approaches to survival models, including those that handle the staggered entry, cannot be applied directly. This paper develops a new general framework to handle such adaptive designs. The new approach is based on marked point processes and differs from existing approaches by considering entry and calender time rather than survival and calender time. Large sample properties, which are essential for statistical inferences, are established. Special attention is given to the Cox model and related score processes. Applications to adaptive and sequential designs are discussed.

Topics In U-Statistics And Risk Estimation

✦Qing Wang, The Pennsylvania State University, Room 325 Thomas Building, University Park, State College, PA 16802 USA, *quw104@psu.edu*; Bruce George Lindsay, Penn State University

Key Words: U-statistics, kernel density estimation, bandwidth selection, risk estimation, unbiased variance estimator, resampling methods

As our motivating problem, we consider U-statistic form estimators for the risk that arises from L2 and Kullback-Leibler loss functions in the context of nonparametric kernel density estimation. These risk estimators can then be used to select the bandwidth that has the smallest risk estimate. In this context, we are interested in how we could estimate the variance of a general U-statistic when it is used as an unbiased estimator of the parameter of interest theta=E(K) where K is a symmetric function of size m. Long established results demonstrate the asymptotic normality of U-statistics and their asymptotic variance under regularity conditions. However, these asymptotic results are not so reliable when the sample size n is not large or the kernel size m is not negligible compared with n. We consider an alternative approach to estimate the variance with a relatively simple form. This variance estimator is the best unbiased and therefore is applicable even for the cases that m/n is a fixed fraction. In addition, two unbiased resampling schemes have been developed to realize the proposed estimator. We also carried out a simulation comparison with some bootstrap variance estimators.

Two Sample Distribution-Free Inference Based On Partially Rank Ordered Set Samples

✦Jinguo Gao , The Ohio State University, Department of Statistics, 1958 Neil Avenue, Columbus, OH 43210, *gao.95@osu.edu*; Omer Ozturk, The Ohio State University

Key Words: Imperfect ranking, Ranking models, Judgment subsets, Rank-sum-test, Pittman efficacy, Ranked set sampling

This paper develops distribution free inference for a location shift model based on a special sampling design. The new sampling design, in pincipal, is similar to a ranked set sampling with some clear di?erences. In the construction of the sample, a small set of experimental units are judgment ranked without measurement by allowing ties whenever the units can not be ranked with high con?dence. These tied units are replaced in judgment subsets. The fully measured units are then selected from these partially ordered judgment subsets. Based on this sampling design, we construct an estimator, a test and a con?dence interval for the location shift parameter. It is shown that the new sampling design is robust against any possible ranking error and has higher e?ciency than its competitor designs in the literature.

Generalized Varying Coeeficient Models: A Smooth Variable Selection Technique

✦Anneleen Verhasselt, Katholieke Universiteit Leuven, Department of Mathematics, Celestijnenlaan 200 B box 2400, Heverlee, 3001 Belgium

We consider nonparametric smoothing and variable selection in generalized varying coecient models. Generalized varying coecient models are commonly used for analyzing the time-dependent eects of covariates on responses, which are not necessary continuous, but for example counts or categories. We present the P-spline estimator in this context and show its estimation consistency for a diverging number of knots, by using an approximation of the link function. The combination of P-splines with nonnegative garrote (which is a variable selection method) leads to good estimation and variable selection. The method is illustrated with a simulation study and a real data example.

334 Reliability and Quantification of Margins and Uncertainties with application to National Security ■

Section on Physical and Engineering Sciences, Section on Quality and Productivity, Section on Statistics in Defense and National Security, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Semi-Markov Processes Toward Applications in National Security

✦Richard Warr, Air Force Institute of Technology, , *richard.warr@afit.edu*

Key Words: Semi-Markov, Statistical Flowgraph, Bayesian Statistics, Laplace Transform, National Security

Multistate models possess an intuitive approach to modeling complex systems. A common simplifying assumption used in multistate models is the semi-Markov property, this assumes that transitions are conditionally independent and readily allows the expression of a likelihood function. The limiting or asymptotic properties of semi-Markov models are well known and widely used, however, their interim properties, or properties before the limiting behavior is accurate, are seldom considered; in-fact it is a common misconception that solutions to these interim properties are not feasible. We demonstrate that solutions are possible and methods to calculate them, specifically, how to compute the state probabilities, first passage distributions, expected number of visits to a state, and others quantities of interest. Additionally, we advocate the use of semi-Markov models in national security problems such as reliability of weapon systems, cost estimation, and projecting personnel requirements. A notional example of projecting the expected costs of the U. S. Government's national threat level is presented.

Reliability Modeling of Nuclear Power Plant Subsystems Using Statistical Flowgraphs

✦David H Collins, Los Alamos National Laboratory, PO Box 1663, MS-F600, Los Alamos, NM 87545, *dcollins@lanl.gov*

Key Words: Nuclear energy, Flowgraphs, Semi-Markov process, Reliability model

Nuclear power plants (NPPs) play an important role in energy security and reduction of airborne pollutants. As part of the U.S. Department of Energy's reactor sustainability program, we are developing models to characterize the reliability and safety of NPP piping subsystems. Subsystems are represented as statistical flowgraphs, with vertices representing states of partial or complete failure and edges representing probability distributions for transitions between states. Failure transitions are driven by processes based on material properties of the pipes, and the physical and chemical dynamics of the fluid being carried. Repair transitions are based on detection of leakage by visual inspection, or non-visible flaws by radiography or ultrasound. Given the complexity of the transition processes, we model the subsystems as semi-Markov processes, using the flowgraph methodology to solve for quantities of interest such as the hazard rate for pipe rupture.

Use of Power Calculations for Gauging Test Adequacy for One-Shot Devices

✦Alix Robertson, Lawrence Livermore National Laboratory, Livermore, CA , *robertson18@llnl.gov*; Rene Bierbaum, Sandia National Laboratory

Key Words: power, monitoring, reliability, one shot

Predicting the performance of one-shot devices with long periods of dormant storage poses unique challenges to the analyst. In this paper, we describe a model of performance change during storage and the implications of this model for life cycle evaluation. The performance model includes different classes of unknown temporal behaviors that must be considered when designing a monitoring program. We discuss the use of power calculations to design and evaluate different monitoring programs in the face of unknown temporal behavior.

Quantifying Reliability Uncertainty From Catastrophic And Margin Defects: A Proof Of Concept

✦John Lorio, Sandia National Labs, , *jflorio@sandia.gov*

Key Words: Method of Moments, Bayesian Analysis, Bootstrap Analysis, System Reliability, Catastrophic Failure Modes, Margin Failure Modes

We aim to analyze the use of component level reliability data, including both catastrophic failures and margin failures, to estimate system level reliability and uncertainty. In this paper, a catastrophic failure is the failure of a component to produce any output and a margin failure is the failure of a component's output to meet a functional requirement.

While much work has been done to analyze margins and uncertainties at the component level, a gap exists in relating this component level analysis to the system level. We apply methodologies for aggregating uncertainty from component level data to quantify overall system uncertainty. We explore three approaches towards this goal, the Classical Method of Moments, Bayesian, and Bootstrap methods. These three approaches are used to quantify the uncertainty in reliability for a system of mixed series and parallel components for which both discrete (pass/fail) and continuous margin data are available. This paper provides proof of concept that uncertainty quantification methods can be constructed and applied to system reliability problems. We also show that the three fundamentally different approaches give comparable results.

Negative Log-Gamma Modeling For Reliability Trends In Series Systems

✦Roger Zoh, Iowa State University, 3410 Snedecor Hall, Department of Statistics, Ames, IA 50011, *rszoh8@gmail.com*; Alyson Wilson, Iowa State University; Scott Vander Wiel, Los Alamos National Laboratory; Earl Lawrence, Los Alamos National Laboratory

Key Words: negative log-gamma, system reliability, prior distribution, trend

Modeling system reliability over time when binary data are collected both at the system and component level has been the subject of many papers. In a series system, it is often assumed that component reliability is linear in time through some link function. Often little or no prior information exists on the parameters of the linear regression, and in a Bayesian analysis they are modeled using very diffuse priors. This can have unintended consequences for the analysis, specifically in the prediction of system reliability. In this work, we consider negative log-gamma distributions as means of specifying prior information on reliability. We first show how our method can be implemented in modeling the reliability of a series system at a given time and then extend to the case where we are interested in modeling reliability over time.

335 Advances in R Software ■

Section on Statistical Graphics, Section for Statistical Programmers and Analysts, Section on Statistical Computing

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

R Package Mvr For Joint Adaptive Mean-Variance Regularization And Variance Stabilization

✦Jean-Eudes J Dazard, Case Western Reserve University, 10900 Euclid Avenue, Center for Proteomics and Bioinformatics, Cleveland, OH 44106, *jxd101@case.edu*; Hua Xu, Case Western Reserve University; J. Sunil Rao, University of Miami

Key Words: R package, Parallel Programming, Normalization and Variance Stabilization, Regularization, Regularized Test-statistics, High-Dimensional Data

We present an implementation in the R language for statistical computing of our recent non-parametric joint adaptive mean-variance regularization and variance stabilization procedure. The method is specifically suited for handling difficult problems posed by high-dimensional multivariate datasets (p >> n paradigm), such as in 'omics'-type data, among which are that the variance is often a function of the mean, variable-specific estimators of variances are not reliable, and tests statistics have low powers due to a lack of degrees of freedom. The R package offers a complete implementation including: (i) normalization and/or variance stabilization function, (ii) computation of mvr-regularized t- and F-statistics, (iii) generation of diverse diagnostic plots, (iv) option of parallel computation, (v) documentation and illustration by real "omics" datasets with FDR-controlled tests results. To make each feature as user-friendly as possible, only one wrapper subroutine per functionality is to be handled by the end-user. It is available as an R package, called 'mean-variance regularization' ('mvr'), downloadable from the CRAN website.

Interactive Parallel Coordinates Plot Based On Qt

✦Yihui Xie, Department of Statistics, Iowa State University, 102 Snedecor Hall, Ames, IA 50011, *xie@iastate.edu*

Key Words: interactive graphics, parallel coordinates plot, Qt, cranvas

The parallel coordinates (par-coords) plot is a common way to visualize high-dimensional data. There have been a number of dynamic statistical graphics systems such as GGobi and Mondrian, in which we can operate par-coords plots interactively. In this paper, we introduce the par-coords plots in a new graphics package cranvas; it is based on the R packages qtbase and qtpaint, which are API's to Nokia's Qt toolkit through R. The par-coords plot in cranvas has both traditional and novel features; for example, we can draw the par-coords plot on top of the corresponding boxplots, manually reorder the axes, and brush the lines in real time, besides, we also have built-in capabilities to impute missing values, automatically reorder the axes by multi-dimensional scaling or ANOVA, and center the axes by the means or medians, etc. Common brushing modes like union, intersection and complement enable us to conveniently query the data using plots. Finally we will give examples based on the data of NRC rankings for statistics departments in the United States.

Visnab: A Interactive Toolkit For Visualizing And Exploring Genomic Data.

✦Tengfei Yin, Iowa STate University, , *yintengfei@gmail.com*; Michael Lawrence, Genentech Research and Early Development; Nicholas Lewin-Koh, Genentech; Heike Hofmann, Iowa State University; Dianne Cook, Iowa State University; Robert Gentleman, Genentech Research and Early Development

Key Words: interactive graphics, next-generation sequencing, genome visualization, statistical cues, massive data

VisNAB (VisNAB is Not A Browser) is a package developed in R and is designed for plotting genomic data. It currently focuses on high-throughput sequencing data. In addition to genome views, VisNAB offers views of other attributes in the data.The genome can be viewed as a track, a Birds-Eye overview, or a circular view. Because it is written in R, it also provides analytical tools for guiding scientists to the information of interest, with all the strength of Bioconductor methods behind it. It is based on the QT library through the packages qtbase and qtpaint, so it's possible to examine huge amounts of data, and in-

teract with the plots. This toolkit, along with other packages in Bioconductor, forms a high performance analytical pipeline for exploring genomic data.

Interactive Maps For Data Exploration In R

✦Heike Hofmann, Iowa State University, , *hofmann@iastate.edu*

Key Words: interactive, graphics, R package, cranvas

Graphics make it possible to place data in their natural framework - for geographical data, maps are particular effective tools to connect the audience to the background information. The cranvas package provides additional interactive tools for working with geographic information in R.

Integration Of Interactive Genome Views With R

✦Michael Lawrence, Genentech Research and Early Development, 94070, *michafla@gene.com*; Peter Danenberg, University of Southern California; Robert Gentleman, Genentech Research and Early Development; Nicholas Lewin-Koh, Genentech

Key Words: interactive graphics, genome browser, R

We have integrated the R platform for statistical computing with IGB, a widely used genome browser written in Java. The goal is to leverage the synergy of the statistical functionality in R, including numerical algorithms and graphics, with the interactive, genome-oriented views provided by IGB. We will demonstrate the control of IGB from R, as well as the use of IGB plugins that, unbeknownst to the naive user, are implemented in R. IGB and R share data structures, without inefficient duplication. R plots, including those generated by the VisNAB package, are coordinated with IGB views. For example, selection in one results in a selection in another. The results of R-based analyses may be displayed in IGB, helping the user quickly locate the interesting parts of a genome-wide dataset. Interesting features discovered in IGB may be passed back to R for more focused analysis. We will provide an overview of these features and give a live demonstration of the software.

336 Statistical Issues in Influenza Surveillance

Section on Statistics in Defense and National Security, Section on Government Statistics, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

2009 Pandemic And Seasonal (1999-2008) Influenza-Associated Outpatient Visits Based On Pneumonia- And Influenza-Coded Data From Vaccine Safety Datalink (Vsd)

✦Hong Zhou, CDC/AREF, , *fwd6@cdc.gov*; JuFu Chen, CDC/Battelle; Paul Gargiullo, CDC; Po-Yung Cheng, Centers for Disease Control and Prevention; James Baggs, CDC; Amber Bishop, CDC/Battelle; Edward Belongia, MFC; Roger Baxter, KPNC; Jason Glanz, Institute for Health Research, Kaiser Permanente Colorado; Allison Naleway, KPNW; Lisa Jackson, GHC; Steven J Jacobsen, Kaiser Permanente Southern California Medical Group; David Shay, CDC

Key Words: influenza, pandemic, pneumonia and influenza, outpatient, modeling

Site- and age-specific ICD-9 coded pneumonia and influenza (P&I) outpatient visits were used in negative binomial regression models which incorporated influenza viral surveillance data as covariates. The influenza-associated P&I outpatient visits were the sum of estimated influenza-associated pneumonia outpatient visits and ICD-coded influenza outpatient visits. The average of annual estimated seasonal influenza-associated P&I outpatient visit rate for all ages was 64 (95% CI: 53-115) per 10,000 persons; the highest rates were in children younger than 2 years (134) and aged 2-17 years (102), followed by persons aged 18-49 (58). The estimated rates of 2009 pandemic influenza-associated P&I outpatient visits for all ages was 223 (95% CI: 208-318) per 10,000 persons; the highest rates were in children younger than 2 years (412) and aged 2-17 years (415), followed by persons aged 18-49 (208). Persons aged 65 years or older had the lowest rates per 10,000 persons for both seasonal (40) and pandemic (67) influenza. The 2009 pandemic caused a substantial increase in the incidence of outpatient visits for P&I compared with prior influenza seasons, especially in persons aged < 50 years.

Comparing Various Methods For Sentinel Surveillance Site Placement

✦Geoffrey Fairchild, University of Iowa Computational Epidemiology, 207 Haywood Dr., Iowa City, IA 52245, *gcfairch@gmail.com*; Alberto Segre, University of Iowa Computational Epidemiology; Philip Polgreen, University of Iowa Computational Epidemiology; Gerard Rushton, University of Iowa Computational Epidemiology

Key Words: influenza, surveillance, geography, algorithm, sentinel, disease

Background: Influenza-like illness data is collected via an influenza sentinel surveillance network at the state level. In order to gather the best disease spread statistics, placement of these sites is important. Methods: We implemented two surveillance site placement algorithms; one maximizes population coverage while the other minimizes average distance. Using influenza data culled from Medicaid billing data, we evaluated the effectiveness of our algorithms compared to the sites hand-picked by the Iowa Department of Public Health (IDPH). We analyze different subsets of ICD-9 codes to determine which are necessary for peak disease detection. We also analyze the temporal aspects of disease spread to determine if a dynamically changing surveillance system would prove beneficial. Results: Simulating the spread of influenza across the state of Iowa, we show that our sites chosen algorithmically outperform the IDPH's hand-picked sites in terms of number of cases detected. Conclusions: Using our models, we can algorithmically place disease surveillance sites across a region in order to gather the best disease spread statistics.

The Distribute Project, Syndromic Surveillance And The New Paradigm For Public Health Data Sharing And Analysis

✦Marc Paladini, New York City Department of Health and Mental Hygiene, , *mpaladin@health.nyc.gov*

Key Words: syndromic, public health, surveillance, influenza

Syndromic surveillance systems typically use non-diagnostic data (e.g. emergency department (ED) chief complaint) to categorize health care encounters into syndromes (e.g. fever and cough as influenza-like illness (ILI)) and look for patterns and anomalies. Although these systems were first developed as early warning systems to detect bioterrorism, many local and state public health jurisdictions routinely use them to monitor and characterize large scale, seasonal disease trends and have found them especially useful for monitoring influenza incidence. The Distribute system is a communal effort organized by the International Society for Disease Surveillance (ISDS) to share summarized syndromic information from local and state health department systems across jurisdictional boundaries and allow them to compare trends. Data shared are totals of ILI-related and all ED visits by visit date, stratified by age group and by 3-digit zip code. ISDS shares data analyses and visualizations with contributing sites and the public. Initial success has enabled further work allowing sites to address issues such as syndrome standardization, signal response and enhanced statistical analysis.

Generation Of Prediction Intervals To Assess Data Quality In The Distribute System Using Quantile Regression

✦Ian Painter, University of Washington, , *ipainter@u.washington.edu*; Julie Eaton, University of Puget Sound; Debra Revere, University of Washington; Bill Lober, University of Washington; Donald Olson, International Society of Disease Surveillance

Key Words: Syndromic surveillance, Data Quality, Quantile regression

Distribute is a national influenza-like-illness (ILI) surveillance project that integrates data from multiple jurisdictions. Distribute works solely with summarized (aggregated) data. Timeliness of the data varies considerably between sites; for many sites data for each encounter date arrives piecemeal, spread over several days. This spread adds additional noise into the data received by the Distribute system. Systematic differences in the timeliness between sources of data can introduce bias into the indicator of interest, the ILI ratio. Quantile regression using the observed relationship between incomplete and complete data is used to calculate prediction intervals for complete data. Some sites have very narrow prediction intervals that indicate the ILI-ratio calculated from incomplete data approximates the complete data ratio very accurately. Other sites show considerable asymmetry in the prediction intervals that indicate bias in the incomplete data ratios. The prediction intervals can be used either to directly measure the uncertainty in the calculated ratio, or their width can be compared with a threshold to determine if the data quality is acceptable.

Analytic Evaluation Of A Standardization Effort For The Distribute Emergency Department Surveillance Project

✦Howard Burkom, Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, Mailstop 8-224, Laurel, MD 20723, *howard.burkom@jhuapl.edu*; Donald Olson, International Society of Disease Surveillance; Marc Paladini, New York City Department of Health and Mental Hygiene; Atar Baer, Public Health-Seattle & King County; Debra Revere, University of Washington

Key Words: disease surveillance, signal-to-noise, influenza

The Distribute project supports emergency department (ED) surveillance with aggregate data to analyze care-seeking behavior and infection spread. Forty public health sites provide data covering ~60% of US ED visits, each using a local preferred syndrome definition. These definitions vary widely, making analytical comparison difficult. In the ILI-S standardization effort, 6 sites provided 4 years of data using a common syndrome definition comprising 3 components-fever/cough, fever/threat, and flu-defined by a formal code set. Evaluation measures included a signal-to-noise calculation based on epidemic intervals reported by the participating sites. Results of comparative visual and statistical analysis were a) clearer inter-regional comparisons for ILI-S than for Preferred ILI, b) similar differences among age groups, but the noise level in preferred ILI made age-based differences harder to compare, and c) improved signal-to-noise ratio for the ILI-S groupings. Among the components of ILI-S, the fever/throat component proved noisiest. Regions varied widely in the relative component proportions. Findings suggest analytic benefits in addition to clarifying regional comparisons.

337 New Advances in Statistical Genetics/Genomics After GWAS ■●

Section on Statistics in Epidemiology, International Indian Statistical Association, Social Statistics Section

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Statistical Methods For Analysis Of Pooled Sequencing Data

✦Zhi Wei, New Jersey Institute of Technology, GITC 4400, Department of Computer Science, University Heights, Newark, NJ 07102, *zhiwei04@gmail.com*

Key Words: next-generation sequencing, pooled sequencing

Although the development of next-generation sequencing technologies (NGS) in the past few years has made the cost of DNA sequencing plummet dramatically, it is still prohibitively expensive to sequence the complete genomes of hundreds of individuals. Just as in pre-GWAS, re-sequencing small targeted genomic regions is now routinely to harness the massive capacity of NGS. Pooled sequencing is used to identify rare variants in targeted regions of the genome in large populations. However, detection of rare variants from pooled sequencing is more challenging than from individual sequencing. Here we propose a new statistical procedure to detect genomic variants from the output of pooled sequencing data. Simulations and analysis of real data will be used to demonstrate the merits of our approach in comparison with existing methods.

Novel Rank-Based Approaches For Discovery And Replication In Large-Scale Association Studies

✦Dmitri Zaykin, National Institute of Environmental Health Sciences, 111 T.W. Alexander Drive, South Bldg (101); Mail Drop: A3-03, Durham, NC 27709, *zaykind@niehs.nih.gov*

Key Words: multiple testing, false discovery, follow-up studies

Initial findings large-scale association studies provide only tentative evidence of association. Due to uncertainty of the findings, researchers often focus on top ranking variants for replication, instead of considering significance thresholds. The rank-based approach can be used to guide both selection of top ranking variants as well as sample size allocation to discovery and replication stages. This approach utilizes the "ranking probability": chances that at least K true positives will rank among a specific number of top leads. To accommodate linkage disequilibrium (LD), ranking probabilities are usually evaluated via study-specific simulations. This approach has limited utility for design of future studies, because the data required for simulations would have yet to be collected. We derive a simple and highly accurate approximations for ranking probabilities, show that an effective number of tests can be used to accommodate LD, and evaluate consequences of ignoring LD. We illustrate how this approach can be used for planning of discovery and replication.

Detecting Joint Effects Of Multiple Genetic Variants Via Composite Haplotype Approach

✦Chia-Ling Kuo, National Institute of Environmental Health Sciences, 1101 Exchange Place Apt 1426, Durham, NC 27713, *chialing.kuo@gmail.com*; Dmitri Zaykin, National Institute of Environmental Health Sciences

Key Words: genetic association analysis, robust statistic, generalized linear model, score test, trend test, Sime's test

Genome-wide association studies have successfully identified thousands of susceptibility loci from marginal association analysis but these variants can only explain a small amount of trait variation. To study undiscovered genetic and environmental factors responsible for the remaining variation, multi-locus association methods with flexibility to incorporate environmental covariates are needed. We define the composite haplotype as a set of alleles across loci, each contributing one allele. The methods we propose for analyzing composite haplotypes use the techniques of frequency filtering and principal component analysis to reduce dimensionality and utilize the regression F-statistic to test for the association with a binary or continuous trait. We show by simulation that compared to a selection of genotype-based and haplotype-based methods, the sum of individual regression F-statistics maintains robust power for a variety of trait models. Since the method is built in the framework of generalized linear models, it is readily applied for various trait types and able to incorporate environmental covariates.

Comprehending Gene-Based Association Signals: A Penalized Regression Approach For Haplotype-Based Analysis With Application In Pharmacogenetic Studies And Individualized Medicine

✦Megan Lee Koehler, North Carolina State University, 1897 Bellwood Dr, Raleigh, NC 27605, *mlkoehle@ncsu.edu*; Jung-Ying Tzeng, Department of Statistics and Bioinformatics Research Center; Howard D Bondell, North Carolina State University

Key Words: haplotype, penalized regression, pharmacogenetics, individualized medicine

Gene-based/marker-set association analysis has become a promising tool for detecting association signals. It has proven to be a powerful approach for detecting global associations between a set of markers and a phenotype. However, not many tools are available for comprehending the overall signals identified. In this work, we propose a penalized-likelihood method that is able to identify the sources of the overall association signals by studying multi-marker genetic and gene-environment interaction effects. The proposed method also has significant utility in pharmacogenetic research, for which the focus is to evaluate genetic-drug interactions and to obtain information useful in the creation of individualized treatment regimes. Simulation studies show that the proposed penalized method has comparable or more power than the standard approach and maintains control on Type I error rates. In addition, a pharmacogenetic application is provided and highlights the utility of the proposed method.

Pathway-Pdt: A Family-Based Pathway Analysis Method

✦Ren-Hua Chung, Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, PO BOX 19132, MIAMI, FL 33101 USA, *rchung@med.miami.edu*; YoSon Park, Hussman Institute for Human Genomics, University of Miami Miller School of Medicine; Margaret A. Pericak-Vance, Hussman Institute for Human Genomics, University of Miami Miller School of Medicine; Eden R Martin, University of Miami Miller School of Medicine; Richard W. Morris, Hussman Institute for Human Genomics, University of Miami Miller School of Medicine

Key Words: Pathway analysis, PDT, Family-based association

Pathway analysis is useful for identifying the joint effects of genes grouped into biologically-based pathways on disease. Pathway analysis may be more powerful than single-marker association tests to identify variants with modest individual effects on a disease but accumulating across genes in a pathway. The development of pathway analysis methods has focused on using unrelated case-control datasets or p-values from genome-wide association tests. We developed Pathway-PDT, a family-based pathway analysis method and an extension of the Pedigree Disequilibrium Test (PDT) (Martin. AJHG 2000). Pathway-PDT defines a score for each gene based on the most significant PDT statistic from PDT statistics for SNPs within a gene (and 20 kb flanking region). Then the weighted Kolmogorov-Smirnov-like running-sum statistic (Wang. AJHG 2007) is calculated for each pathway. A permutation procedure is used in Pathway-PDT to approximate the distribution of the running-sum statistic. We used simulations to verify that Pathway-PDT maintains correct type I error rates under different scenarios. Our simulation results also suggested that Pathway-PDT can have more power than methods using p-values only.

338 Perspectives on Managing Data Confidentiality and Access Issues by Government, Academic, and Non-profit Resarchers ■

Section on Survey Research Methods, Section on Government Statistics, Section on Health Policy Statistics, Social Statistics Section, Committee on Privacy and Confidentiality, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

The Microdata Analysis System At The U.S. Census Bureau

✦Michael Freiman, U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD , *michael.freiman@census.gov*; Michael Freiman, U.S. Census Bureau; Laura Zayatz, U.S. Census Bureau; Lisa Singh, U.S. Census Bureau; Jiashen You, University of California-Los Angeles; Michael DePersio, U.S. Census Bureau

Key Words: Data Confidentiality, Remote Access Servers, Universe Subsampling, Synthetic Data, Regression

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code, which promises to protect the confidentiality of all survey respondents. Additionally, the Census Bureau has the responsibility to release high quality data products while maintaining this confidentiality. This paper discusses a Microdata Analysis System (MAS) under development at the Census Bureau, which allows users to perform statistical analyses on microdata without having access to the data themselves. The data are also perturbed before analysis, and the results of some analyses may be withheld. Regression and tabulation are currently being implemented, and more options are planned. For ordinary least squares regression, we will give diagnostic plots that mimic residual plots but use only synthetic data. We discuss the reason for developing the MAS and give an overview of the system's capabilities, then describe the rules that protect confidentiality of data. The confidentiality rules implemented so far are generally of two types: restrictions on how a user may define the universe (the full dataset chosen to be analyzed) and restrictions on what regression output may be given.

Research Access To Statewide Longitudinal Data Systems

✦Satkartar Kinney, NISS, , *saki@niss.org*

Key Words: Data Access, Education Data

Statewide Longitudinal Data Systems (SLDS) are intended to enhance the ability of States to efficiently and accurately manage, analyze, and use longitudinal education data, including individual student records. Since 2005, 41 states and the District of Columbia have received grants to develop SLDS, currently in varying stages of completion. State Educational Agencies (SEAs) are extremely diverse in their resources and expertise to conduct research and/or allow external researchers to use their data. Providing external researchers access to SLDS data can be non-trivial and resource intensive for the agency; however, several SEAs and other agencies can and do provide data to external researchers in compliance with confidentiality laws. We discuss approaches and barriers to maximizing the use of SLDS data improve educational outcomes.

The Feasibility Of Using U.S. Census 2000 Public Use Microdata Sample (Pums) To Evaluate Population Uniqueness For Population-Based Cancer Microdata

✦Mandi Yu, National Cancer Institute, 6116 Executive Blvd., Suite 504, Rockville, MD 20852, *yum3@mail.nih.gov*; David Stinchcomb, Westat, Inc.; Kathleen Cronin, National Cancer Institute

Key Words: Cancer Registry Data, Disclosure Risk, Population Unique, Small Area, ACS, PUMS

The cancer registry data collected by NCI's SEER Program has been the most authoritative source for describing cancer incidence and survival. The release of high quality and confidential data is central to the agency's mission. While the internal disclosure threat presented by record uniqueness has been well addressed, little consideration has been given to the external threat in which a data intruder seeks to find out whether a known person has cancer by matching his characteristics with those from registries' records. In this presentation, we develop a non-parametric approach to estimate the proportion of record unique patients who are also population unique. We match key variables between SEER county-level data and Census 2000 PUMS in which county codes are multiply imputed. The methods can also be applied to future assessments in which we use yearly updated ACS PUMS. The results show that the risk estimates tend to be conservative compared with those calculated using 100% Census 2000 summary data. The upward bias is in the neighborhood of 2 to 3 times. The findings will guide disseminating small area SEER data and implementing of confidentiality control procedures.

Simulating Geography For Microdata Disclosure Via Sparse Multinomial Probit Models

✦Lane F Burgette, Duke University, Box 90251, Durham, NC 27708, *lb131@stat.duke.edu*; Jerome P. Reiter, Duke University

Key Words: Confidentiality, Multinomial probit, Potts model, Spatial, Synthetic

Public release of spatially-referenced microdata can entail significant risk that motivated intruders will be able to learn the identities of respondents who provide sensitive data. To mitigate this risk, it is standard to aggregate data over large geographic areas, which can degrade the utility of the data for legitimate researchers. As an alternative, we propose methods to produce synthetic sets of areal identifiers. Our goal is to simulate multiple sets of data that--on average--retain the statistical properties of the observed data, while protecting respondents' anonymity. We propose methods to simulate areal identifiers using a multinomial probit model. Because this results in a model that (in typical applications) will have hundreds or even thousands of response categories, we propose a sparse structure for the multinomial model. Further, we suggest a simplified, latent Potts model structure for the regression coefficients, which can help to preserve spatial relationships. We demonstrate our methods on simulated and genuine data.

Applicability of Basic Separability Principles to Enhance the Operational Efficiency of Synthetic Tabular Data Generation Procedures in Multidimensional Table Structures

✦Ramesh Atmaram Dandekar, U S Department of Energy, EI-20, 1000 Independence Av, Washington, DC 20585, *Ramesh.dandekar@gmail.com*

Key Words: Statistical Disclosure, perturbation, tabular data, establishment data, individual data, privacy

Hyperlinked copy of this abstract is available at URL http://mysite.verizon.net/vze7w8vk/AbstractJSM2011.pdf Dandekar2001 proposes using synthetic tabular data generation (i.e. controlled adjustments to

tabular data - CTA) as an alternative to complementary cell suppression procedures. Dandekar2009 addresses quality aspects of CTA protected tabular data with an objective to completely replace conventional complementary cell suppression procedures with a new tabular data protection method. The proposed method is a hybrid of three different tabular data protection methods. CTA has already been demonstrated to be equally effective on multi-dimensional counts data and magnitude data containing complex hierarchies and linked table structures. In this paper we go a step further to demonstrate how the operational efficiency of the CTA procedures could be enhanced significantly by applying basic separability principles to complex table structures. The proposed enhancements have significant potential to cut down on computational resources by reducing the problem size in terms of the number of variables and associated mathematical constraints.

339 Statistical Considerations in the Analysis of Progressive Free Survival

Biopharmaceutical Section, Section for Statistical Programmers and Analysts, Section on Statistics in Epidemiology

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Predicting Analysis Time In Event-Driven Clinical Trials With Event-Reporting Lag

✦Jianming Wang, Novartis Oncology, One Health Plaza, East Hanover, NJ 07936, *jianming.wang@novartis.com*; Chunlei Ke, Amgen; Qi Jiang, Amgen Inc.; Charlie Zhang, Exelixis; Steve Snapinn, Amgen, Inc.

Key Words: Event-Driven, Clinical Trial, Analysis Time Prediction, Bayesian

For a clinical trial with a time-to-event primary endpoint, the accrual of the events of interest during the course of the clinical trial may determine the timing of the analysis and provide useful information for the planning of resources and drug development strategies. It is of interest to predict the analysis time early and accurately. Currently available methods use either parametric or non-parametric methods to predict the analysis time based on accumulating information about enrollment, event and drop-out rates. However, these methods are usually based on the implicit assumption that the available data are complete at the time of performing prediction. This may not be true when it takes an uncertain amount of time to report an event to the database. As a consequence, the data will be incomplete at the time of performing prediction. Specifically, some patients without a reported event might have had an event that had not yet been reported. Ignoring this event-reporting lag could substantially impact the accuracy of the prediction. In this talk, we describe a general parametric approach to predict analysis time by incorporating event-reporting lag using a Bayesian method.

The Effect Of Periodic Follow-Up: Design And Analysis Of Progression-Free Survival Data

✦Zhigang Zhang, MSKCC, 307 E. 63rd Street, 3rd Floor, New York, NY 10065, *zhangz@mskcc.org*

Key Words: Clinical Trials, Progression-Free Survival, Interval Censoring

In clinical trials, progression-free survival (PFS) is often used as an endpoint to assess the efficacy of a therapy. In contrast to overall survival (OS), it is not affected by the second-line therapy and can be observed earlier. However, the measurement of PFS can be difficult both clinically and statistically. As a matter of fact, its accuracy highly depends on interpretation from the radiologists and timing of the scheduled follow-up examinations. The latter is known as interval censoring in biostatistics. In this talk, I will focus on this aspect and discuss issues in experimental design and statistical analysis of PFS. I will present some simulation studies and real data analysis results to illustrate my points.

Some Practical Issues In Pfs Analysis From Industry Perspective

Jingyuan Wang, Millennium Pharmaceuticals; Connie Lee, Millennium Pharmaceuticals; Jing Xu, Millennium Pharmaceuticals; ✦Yuanjun Shi, Millennium Pharmaceuticals, , *Yuanjun.Shi@MPI.com*

Key Words: PFS, censoring, interval censored

We will talk on a few practical issues in PFS analyses. Monte Carlo simulations are used to study the impact of various methods on events time and censoring handling, including unscheduled assessments, frequency of assessments, start of new therapies etc. Methods recommended by regulatory agencies will be discussed. Theoretical explanation may be provided to support some of the simulation results.

Regulatory Considerations In The Evaluation Of Progression-Free Survival

✦Rajeshwari Sridhara, USFDA, 10903 New Hampshive Ave., WO 21, Rm 3512, Silver Spring, MD 20993, *rajeshwari.sridhara@fda.hhs.gov*

Key Words: PFS, Censoring rules

Improvement in overall survival and patient reported outcomes provide direct clinical benefit to patients. However, consideration of progression-free survival (PFS) as the primary endpoint for demonstration of efficacy for approval of drug products is subject to demonstration of statistically persuasive effect & a magnitude of effect that will translate to a meaningful clinical benefit with an acceptable risk-benefit profile. In open label studies independent radiologic review is recommended as progression assessments are complex and subjective. This can result in informed censoring when the investigator and independent reviewers do not agree on the progression assessment. While we acknowledge that any choice of censoring rule in evaluating PFS has problems including informed censoring, the current FDA recommendation for the PFS analysis is that the PFS data be censored on the date of the last tumor assessment documenting absence of progression for patients who have no documented progression, or change therapy, or discontinue due to toxicity or personal preference. FDA's rationale for this current recommendation and ongoing research will be presented.

Issues In The Analysis Of Progression-Free Survival From A Cancer Clinical Trial

✦Dianne M Finkelstein, Mass General Hospital/Harvard SPH, 50 Staniford Street, Suite 560 (MGH Biostatistics Center), Boston, MA 02114, *dfinkelstein@partners.org*

Key Words: clincial trial, interval censoring, dependent censoring, PFS, cancer

The use of progression free survival (PFS) as a primary endpoint in cancer trials must consider several issues to ensure validity of this outcome as a surrogate for survival. First, although a trial is designed to evaluate progression at regular prescribed time points, recurrence can be recorded at times outside these times because visits are missed, resulting in interval censored data. Second, the time that disease is evaluated may be determined by a report of symptoms, which could result in a dependent censoring pattern. Third, patients may stop or change therapy or withdraw from the study, which could result in a loss of power or a bias in the analysis. We will discuss these issues and methodology that can be used to appropriately handle these various censoring issues in the analysis of PFS in cancer clinical trials.

340 Bayesian Nonparametric and Semiparametric Methods 1 ●

Section on Bayesian Statistical Science, International Indian Statistical Association, Section on Nonparametric Statistics

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Bayesian Clustering Of Curves And The Search Of The Partition Space

✦Silvia Liverani, University of Bristol, School of Mathematics, University Walk, Bristol, BS8 1TW UK, *s.liverani@bristol.ac.uk*

Key Words: Bayesian statistics, Bayes factors, clustering

This thesis is concerned with the study of a Bayesian clustering algorithm, proposed by Heard et al. (2006), used successfully for microarray experiments over time, and it focuses not only on the development of new ways of setting hyperparameters so that inferences both re?ect the scienti?c needs and contribute to the inferential stability of the search, but also on the design of new fast algorithms for the search over the partition space. First we use the explicit forms of the associated Bayes factors to demonstrate that such methods can be unstable under common settings of the associated hyperparameters. We then prove that the regions of instability can be removed by setting the hyperparameters in an unconventional way. Moreover, we demonstrate that MAP (maximum a posteriori) search is satis?ed when a utility function is de?ned according to the scienti?c interest of the clusters. We then focus on the search over the partition space. In model-based clustering a comprehensive search for the highest scoring partition is usually impossible, so here we propose to view clusterings as elements of the lattice of partitions as well as encoding clustering as a weighted MAX-SAT problem.

Modelling Via Normalisation For Parametric And Nonparametric Inference (Submitted For Savage Award - Theory And Methods)

✦Michalis Kolossiatis, Department of Commerce, Finance and Shipping, Cyprus University of Technology, Limassol, 3036 Cyprus, *michalis.kolossiatis@cut.ac.cy*; Jim E. Griffin, School of Mathematics, Statistics and Actuarial Science, University of Kent at Canterbury ; Mark F J Steel, Warwick University

Key Words: Bayesian nonparametrics, Normalized random measures, Markov chain Monte Carlo, Dependent distributions, Overdispersion

Bayesian nonparametric modeling has recently attracted a lot of attention, mainly due to the advancement of various simulation techniques, especially Monte Carlo Markov Chain (MCMC) methods. We propose some Bayesian nonparametric models for grouped data, which make use of dependent random probability measures. These probability measures are constructed by normalizing infinitely divisible probability measures (for example, gamma processes) and exhibit nice theoretical properties. Implementing these models is easy, using mainly MCMC methods. We also add a split-merge step in these algorithms, in order to improve mixing. The proposed models are then embedded in a stochastic frontier setting, in order to study the efficiency of some hospital firms, based on their ownership status and staff-to-patients ratio. We also propose a new, n-dimensional distribution on the unit simplex, that contains many known distributions as special cases. The univariate version of this distribution is used as the underlying distribution for modeling binomial probabilities. Using simulated and real data, it is shown that this proposed model is particularly successful in modeling overdispersed count data.

Semiparametric Stochastic Modeling Of The Rate Function In Longitudinal Studies(Applications)

✦Bin Zhu, Duke University, Department of Statistical Science, and Center for Human Genetics, , *bin.zhu@duke.edu*; Peter Song, University of Michigan; Jeremy Michael George Taylor, University of Michigan

Key Words: Euler approximation, Functional data analysis, Gaussian process, Rate function, Stochastic differential equations, Semiparametric stochastic velocity model

In longitudinal biomedical studies, there is often interest in the rate functions, which describe the functional rates of change of biomarker profiles. This paper proposes a semiparametric approach to model those functions as the realizations of stochastic processes, using the stochastic differential equations. These processes are dependent on the covariates of interest, and are expected to be centered on some parametric forms while allowing significant deviations from these functional expectations. An efficient Markov chain Monte Carlo algorithm is developed for the inference. The proposed method is compared with several existing methods in aspects of goodness-of-fit and more importantly the ability of forecasting via validation functional data in a simulation study and an application to prostate-specific antigen profiles.

341 New Advances in Statistical Learning & Data Mining

Section on Statistical Learning and Data Mining, Section for Statistical Programmers and Analysts

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Probabilistic Biclustering

✦Thao Duong, University of California, Irvine, , *ttduong@uci.edu*; Hal Stern, Department of Statistics; Irvine Curtis Deutsch, University of Massachusetts Medical School

Key Words: data mining, biclustering, algorithms, bayesian, mcmc, mixture model

Biclustering refers to identification of subsets of individuals (units) and subsets of measurements (features) that define interesting partitions of the data. We formulate a parameter probability model-based biclustering approach. We handle the problem of identifying the discriminative features and the observations in each cluster by introducing binary latent vectors (for rows and columns seperately) for each cluster. The resulting method selects the optimal number of clusters, discriminating features and observations in each group simultaneously. We apply the method to simulated and real examples.

The Relative Power Of A Support Vector Regression Approach To Survival Analysis And The Cox Proportional Hazards Model

✦Douglas A. Powell, Aureon Biosciences, Inc, , *douglas.powell@aureon.com*; Faisal M. Khan, Aureon Biosciences

Key Words: Statistical Power, Survival Analysis, Support Vector Machines, Cox Proportiona Hazards Model

The ability to distinguish between high and low risk is critical for survival models. The statistical power of the standard Cox model is known to be sensitive to the sample size, model size, and measurement error (noise) in the features and the event rate. We compared the Cox Model against a new Support Vector Regression for Censored Data (SVRc) approach to survival analysis. Data were simulated varying sample size (4 levels), model size (4 levels), feature noise (3 levels) and event rate (3 levels). Power was assessed by the model's hazard ratio (HR) and the proportion of statistically significant HRs. Results indicate that SVRc was superior to the Cox Model for both criteria in most of the 144 combinations. The HR improved by a median rate of 9.6% with over 25% of the 144 conditions manifesting more than a 40% increase. The SVRc empirical power matched or exceeded the Cox model's in 84% of conditions, with an increase of 24% or more for 30% of the conditions. The absolute worst performance of SVRc was a 12% decline in one experiment. The results suggest that SVRc improves predictive power compared to the Cox model and can attain a given power with smaller sample sizes.

Multi-Objective Forecast Combination

✦Gang Cheng, University of Minnesota at twin cities, School of Statistics, 313 Ford Hall, 224 Church Street S.E., Minneapolis, MN 55455, *chen2285@umn.edu*; Yuhong Yang, University of Minnesota, Twin Cities

Key Words: multi-objective, forecast combination, adaptive loss function, forecast evaluation

Forecast is commonly used in many fields such as economics and finance to serve a specific goal or multiple goals simultaneously. In the literature of forecast combinations, there have been only a few studies exploring how to serve multiple goals simultaneously in a single forecast combing process. In this work, we propose a multi-objective combing method, which build on the AFTER algorithm (Yang, 2004a) by using an adaptive integrated loss function to serve multiple goals. The theoretical results for many versions of this method are obtained, and they suggest that the combined forecasts automatically and consistently perform as well as the best individual among the pool of forecast candidates under each of the multiple evaluation criterions interested. Simulations and data examples show that the proposed method outperforms AFTER when multiple evaluation criterions are of similar importance and performs similarly if one of the criterions is dominant. Besides, the simulation shows the potential advantages of the multi-objective AFTER algorithm comparing with some commonly used combing methods.

Assessing Data Regularity In Complex Wavelet Domain: Application To Mammography Image Classification

✦Seonghye Jeon, Georgia Institute of Technology, 765 Ferst Drive NW, Atlanta, GA 30332, *sjeon8@mail.gatech.edu*; Seonghye Jeon, Georgia Institute of Technology; Brani Vidakovic, Georgia Institute of Technology

Key Words: image classification, regularity index, complex wavelets, mammography image

A wide range of complex structures in nature exhibits irregular behavior in both time and scale. Although irregular, the phenomena can be well modeled by multifractal processes that are quantified by statistical similarity of patterns at different scales. Wavelet transform is a powerful tool for analyzing the complex structures of the data and assessing the regularity of the multifractal processes. Complex wavelets have been advocated as solutions to some of the limitations of the real-valued wavelet transforms. Apart from the Haar wavelet, complex wavelets are only compactly supported orthogonal wavelets which are symmetric. Another advantage of the complex wavelets is the complementary phase information that describes the coherent structure of the image. Although complex wavelet transform has been used in various areas, ours are the first to explore the comprehensive regularity indices. We extend the wavelet spectrum and multifractal spectrum to the complex wavelet domain and propose new regularity descriptors including phase information and coherence function. This study is motivated and illustrated by mammography image classification.

A Universal Correlation Coefficient

✦nuo xu, university of alabama at birmingham, , *nuoxu@uab.edu*

Key Words: correlation coefficient, information measure, dependency and association, rank statistics

Developed by Galton and Pearson more than a century ago, the correlation coefficient is still one of the most widely known and used indexes in statistical analysis. Its susceptibility to outliers and incapability of capturing no-linear relationships were remedied to a certain degree by its many variants developed by statistician such as Spearman and Kendall. However, these indexes are still incommensurate among different type of variables. A universal correlation coefficient is developed based on empirical observation of difference on the rank sequence of original data and its expection. This index can capture any form of relationship and serve as a commensurate correlation measure among pairs of variable of different types. A comparative study is presented to show its superiority over other correlation indexes.

Naturally Efficient Sparsity Tuner For Kernel Regression

✦Ernest Fokoue, Rochester Institute of Technology, 98 Lomb Memorial Drive, Rochester, NY 14623 USA, *ernest.fokoue@rit.edu*

Key Words: Regression, Information Matrix, Sparsity, Structured Prior Matrix, Kernel, Support Points

We propose a novel approach to achieving sparse representation in kernel regression through a straightforward algorithm that consists in a refinement of the maximum a posteriori (MAP) estimator of the weights of the kernel expansion. Our proposed method combines structured prior matrices and functions of the information matrix to zero in on a very sparse representation. We show computationally that our naturally efficient sparsity tuner (NEST) achieves a very sparse and predictively accurate estimator of the underlying function, for a variety of choices of the covariance matrix of our Gaussian prior over the weights of the kernel expansion. Our computational comparisons on both artificial and real examples show that our method compete very well - usually favorably - with the Support Vector Machine, the Relevance Vector Machine and Gaussian Process regressors.

Automated Pattern Recognition For Dynamical Space-Time Systems Based On Local Statistical Complexity

✦Georg M Goerg, Carnegie Mellon University, 5000 Forbes Avenue, 132 Baker Hall, Pittsburgh, PA 15213, *gmg@stat.cmu.edu*

Key Words: optimal prediction, pattern recognition, causality, predictive distributions, complexity, space-time systems

Many methods in statistics, machine learning, and signal processing, such as pattern recognition in videos, try to extract informative structures from a dynamic system and remove noisy uninformative parts. Although such methods work well in practice, they often do so because they have been tuned to work in a very particular setting, and thus may break down when conditions and properties of the data do not hold anymore. It would be very useful to have an automated method showing us informative patterns, which does not rely on any particular model or data structure of the system. Shalizi (2003) showed for discrete fields that an automated pattern discovery can be constructed by a characterization and classification of local conditional predictive distributions. The underlying idea is that statistically optimal predictors not only predict well, but for this very reason also describe the data well, and therefore reveal informative structure inherit in the system. I extend these methods to obtain a fully automated pattern recognition for continuous-valued space-time systems, with applications to simulated as well as real-world spatial dynamics.

342 New Advances in ROC Methodology

Biometrics Section, ENAR, International Indian Statistical Association, Section on Statistics in Epidemiology

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Maximizing The Area Under Roc Curve With Grouped Lasso

✦Sungwoo Choi, University of Maryland, Baltimore County, Department of Mathematics & Statistics, 1000 Hilltop Circle, Baltimore, MD 21250 USA, *schoi8@umbc.edu*; Junyong Park, University of Maryland Baltimore County

Key Words: ROC curve, AUC, Mann Whitney statistic, grouped lasso

In this talk, we present some results on the estimation of AUC (Area Under ROC curve). In multivariate case, linear combination is a typical way to combine multiple variables and it is incorporated with logistic regression or Mann-Whitney statistic as an objective function. We discuss the optimal score function to maximize AUC and estimate it in high dimension based on selecting meaningful variables by grouped lasso technique. Logistic regression with grouped lasso has been popular whereas Mann-Whitney statistic with grouped lasso has not been widely recognized. We demonstrate the performance of Mann-Whitney statistic with grouped lasso and compare it with logistic with lasso and grouped lasso based on simulations and real data examples.

Empirical Likelihood-Based Inferences In Roc Analysis With Covariates

✦Gengsheng Qin, Georgia State University, 30 Pryor Street, Atlanta, GA 30303, *gqin@gsu.edu*; Baoying YANG, Southwest Jiaotong University

Key Words: AUC regression, empirical likelihood, ROC regression, bootstrap, confidence region

In ROC analysis, the area under the ROC curve (AUC) is a popular one number summary index of the discriminatory accuracy of a diagnostic test. Accounting for covariates can improve diagnostic accuracy of the test. Regression models for the ROC curve and the AUC are two means to evaluate the effects of the covariates on the diagnostic accuracy. In this paper, empirical likelihood (EL) methods are proposed for the AUC regression model and the ROC regression model respectively. For both of the regression parameter vectors in the AUC regression model and the ROC regression model, it is shown that the limiting distributions of their EL ratio statistics are the weighted sum of independent chi-square distributions. Confidence regions can be constructed for the parameter vectors in the regression models based on the newly developed empirical likelihood theories. We can also construct confidence interval for the covariate-specific AUC. Simulation studies are conducted to compare the relative performance of the proposed EL-based methods with the existing method in AUC regression. Finally, we illustrate the proposed methods with a real data set.

Receiver Operating Characteristic (Roc) Curve-Based Variable Selection For Logistic Regression Models

✦Jodi Lapidus, Oregon Health & Science University, , *lapidusj@ohsu.edu*; Mara Tableman, Portland State University; Aaron Baraff, Oregon Health & Science University

Key Words: logistic regression, biomarkers, variable selection, ROC curves

Many recent investigations have been dedicated to identifying and evaluating biomarkers, and these studies often focus on combining information from multiple markers to classify disease. While there are ample classification methods proposed in the literature, Pepe (2003) showed that decision rules based on the likelihood ratio function, or equivalently, the risk score, are optimal. Logistic regression can be used to generate a risk score, and the c-statistic or area under the ROC curve (AUROC) can be computed to assess classification performance. When several candidate biomarkers are collected, it is labor-intensive to check performance of all possible combinations. We outline a procedure to select markers for inclusion in a logistic regression model based on improvement in AUROC. We note the equivalence of a nonparametric two-sample test statistic and AUROC, and use this to select the first marker for the model. We make use of the jagged ordered multivariate optimization algorithm for partial ROC curves outlined in Baker (2000). We illustrate our algorithm on various sized datasets, and contrast the results to models fit with standard variable selection methods.

Regression Models For Positive And Negative Predictive Values

✦Michael Sachs, University of Washington, Dept of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, *sachsmc@gmail.com*

Key Words: biomarkers, regression, predictive values, diagnosis

In many clinical studies, it is of interest assess the risk of incurring some binary outcome, such as death or disease, based on current medical test results. This paper focuses on the positive predictive value as a tool to assess the prospective accuracy of a continuous test. In some cases, a covariate may affect the accuracy of a continuous test, or different thresholds for high-risk versus low-risk are needed for different groups defined by some covariate. We present two regression models for the positive predictive value of a continuous test for a binary outcome. The first model directly links the positive predictive value to a linear predictor. The second model is an indirect approach. We present some theoretical results, and then illustrate the use of these models in two different studies. The first is a longitudinal study of depression, in which it is of interest to assess the change in accuracy of a screening test when administered repeatedly over time. The second illustration uses our model to develop education- and race-specific thresholds for the mini mental state examination in predicting clinical diagnosis of Alzheimer's disease.

Equivalence Of Improvement In Area Under Roc Curve And Linear Discriminant Analysis Coefficient Under Assumption Of Normality.

✦Olga Demler, Boston University, 30 Edge Hill Rd, Newton, MA 02467 USA, *demler@bu.edu*; Michael Pencina, Boston University; Ralph B. D'Agostino, Sr, , Boston University

Key Words: AUC, ROC, discrimination, linear discriminant analysis, significance of the predictor, logistic regression

Area under the Receiver Operating Characteristics Curve, (AUC of ROC) is a widely used measure of discrimination in risk prediction models. This study was motivated by numerous reports that often the added predictor is statistically significantly associated with the outcome but fails to produce significant improvement in the AUC. We demonstrate that under the assumption of multivariate normality and employing linear discriminant analysis to construct the risk prediction tool, statistical significance of the new predictor(s) is equivalent to the statistical significance of the increase in AUC. We extend this result for unequal, non-proportional variance-covariance matrices of predictors within cases and non-cases. The result holds asymptotically for logistic regression under assumptions of normality and equal covariance matrices. Our practical example from the Framingham Heart Study data suggests that the finding might be sensitive to the assumption of normality.

The Linear Combinations Of Markers Which Maximize The Partial Area Under The Roc Curves

✦Man-Jen Hsu, National Chengchi University, Taiwan, *95354503@nccu.edu.tw*; Huey-Miin Hsueh, Department of Statistics, National Chengchi University

Key Words: Partial area under curve, pAUC, ROC curve, Sensitivity, Specificity

As biotechnology has remarkable progress nowadays, there is a great improvement in data collecting procedure with lower cost and higher quality. When multiple potential markers are available in constructing a diagnostic tool of a disease, an effective approach is to combine the information to build one single summarizing indicator. For continuous-scaled data, the linear combination is popular due to its easy interpretability. Su and Liu (1993) derived the best linear combination under the criterion of maximal area under the ROC curve. In many investigations, the emphasis is placed only on limited extent, instead of the whole curve. The goal of this study is to find the best linear combination that maximizes the partial area under a ROC curve (pAUC). To find the solution analytically, the first derivative of the pAUC under normal assumption is derived. Because the pAUC maximizer may not be unique in some cases, the existing algorithm is inadequate and thus we propose a revised algorithm by adopting several initial points. Intensive numerical studies are perform and the results justify the adequacy of the proposed algorithm. Real examples are also provided for illustration.

343 Flexible Parametric Modeling in Survival Analysis

Biometrics Section, Section on Statistics in Epidemiology
Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Regression Models With Covariates Missing In Nonmonotonic Patterns

✦Yang Zhao, University of Regina, Department of Mathematics and Statistics, College West 307.14, Regina, SK S4S 0A2 Canada, *zhaoyang@uregina.ca*; Meng Liu, University of Regina

Key Words: Complete-case analysis, Estimating equation, Nonmonotonic missing patterns, Parametric working model, Weighted complete-case analysis

This research proposes estimation methods for regression models with covariances missing in arbitrary nonmonotonic patterns. It explains the idea of using a sequence of parametric working models to extract information from incomplete observations and computing efficient estimates of regression parameters. The proposal is general and can be applied to analysis for regression models with arbitrary nonmonotonic missing data patterns. We use simulation studies and an analysis of a real data example to illustrate the proposed method.

Estimating Incidence Rate On Current Status Data Imputing Partial Status With Application To Pediatric Cancer Survivor Study

Shesh N. Rai, University of Louisville; ✦Jianmin Pan, University of Louisville, , *jianmin.pan@louisville.edu*; Xiaobin Yuan, University of Louisville

Key Words: Phase IV clinical trial, Cardiotoxicity, Cross-section survey data, Interval censored data, K-M method, Imputation

New drug discovery has dramatically improved survival, but with long-term adverse events. This motivates the examination of adverse outcomes such as long-term drug toxicity in a Phase IV trial. An ideal approach to monitor long-term toxicity is to systematically follow the survivors, which is generally not feasible. Instead, cross-sectional surveys are sometimes conducted to investigate the long-term effects of cancer treatment. In these studies, one objective of interest is to estimate the cumulative incidence rates of toxicity with specific interest in fixed-term rates. However, such data poses many issues: incomplete data, competing risks and selection bias, etc. An example for this study is described in Hudson et al. (2007) studying the effect of cardiotoxicity due to anthracyclines exposure during cancer treatment. The main issue in such studies is the high missing rate. In this paper, we impute the missing values using multivariate regression method under some parametric assumptions, combine with methods previously described to estimate the cumulative incidence rates in an illness-death/failure model, and compare with the results obtained without using imputation.

Likelihood Ratio Tests For A Change Point In Weibull Hazard Models With Covariate Information

✦Matthew Richard Williams, Department of Statistics, Virginia Tech, 403 F Hutcheson Hall, Blacksburg, VA 24061, *mrwilli@vt.edu*; Dong-Yun Kim, Virginia Tech

Key Words: change point, Donsker class, hazard function, likelihood ratio test, local asymptotic normality, Weibull

We consider likelihood ratio tests for the presence of an unknown change point in a Weibull Hazard model with known shape parameter. We utilize local asymptotic normality to derive the limiting process for the test. With this new framework, we are able to extend previous results to include covariate information. We demonstrate our method for the case of staggered-entry and type I censoring, and provide simulation results comparing our method to some of the available alternatives.

Robust Working Models In Survival Analysis Of Randomized Trials

✦Jane Paik, Stanford University, 1070 Arastradero Road, Palo Alto, CA 94304 United States, *janepaik@stanford.edu*

Key Words: Model misspecification, Cox proportional hazards model, parametric proportional hazards model, partial likelihood, hypothesis testing, randomized trials

An important outcome in randomized clinical trials is the time from the randomization to a treatment until a failure. The hypothesis we are interested in testing is whether the treatment has an effect on the survivor function conditional on strata of baseline variables. We identifiy survival regression models that give rise to hypothesis tests with asymptotically valid type I error under the null hypothesis, regardless of model misspecification. Using a direct application of Lemma 2 from Rosenblum and van der Laan (2009), we demonstrate that parametric and Cox proportional hazards models, as well as multiplicative hazards models with certain non-exponential link functions are robust to misspecification when the censoring distribution does not depend on the treatment given covariates. We will also demonstrate that a class of adjusted parametric models is robust to misspecification even when the censoring distribution depends on the treatment assignment.

Joint Parametric Modeling Of Overall Survival And Progression-Free Survival

✦Somesh Chattopadhyay, US Food and Drug Administration, 10903 New Hampshire Avenue, Bldg 21, Rm 3525, Silver Spring, MD 20993-0002, *somesh.chattopadhyay@fda.hhs.gov*; Shenghui Tang, US Food and Drug Administration; Rajeshwari Sridhara, USFDA

Key Words: Overall survival, Progression-free survival, Joint modeling

Overall survival (OS), defined as the time from randomization to death due to any cause, and progression-free survival (PFS), defined as the time from randomization to either documented disease progression or death from any cause (whichever occurs earlier), are two important endpoints in clinical trials commonly used in oncology. PFS is sometimes used as a surrogate reasonably likely to predict for OS. It is important to find out how these two endpoints are related. In this research we model these two endpoints jointly using parametric models. We explore the properties of the joint model and how to predict OS from PFS using the model.

Parametric Regression Models For Cumulative Incidence Functions Based On Generalized Odds-Rate Models And Modified Logistic Functions

✦Haiwen Shi, University of Pittsburgh, Department of Statistics, 2717 Cathedral of Learning, Pittsburgh, PA 15260, *has9@pitt.edu*; Yu Cheng, University of Pittsburgh

Key Words: competing risk, cause-specific hazard function, cumulative incidence function, long-term incidence, modified three-parameter logistic mode, parametric modeling

We propose a new regression model for cumulative incidence functions (CIFs) in competing risk data. Fine and Gray (1999) developed a semiparametric regression model for CIFs and Jeong and Fine (2007) extended a generalized odds-rate model to the competing risk setting with Gompertz models as the baseline CIFs. We propose to use a modified logistic function (Cheng, 2009) for the baseline CIFs in the parametric modeling. The inference procedure relies on standard maximum likelihood methods. Extensive simulations are run to evaluate the performance of the coefficient estimators from our proposed model, the Gompertz model and the semiparametric model. We also apply the three models to Cache County Study data examining the effects of covariates on the CIF of dementia. Our numerical results suggest that our parametric model be more flexible than the Gompertz model and have comparative performance with the semiparametric model for the primary cause. In addition, our parametric model considers the fact that the cumulative probabilities from different causes should eventually add up to one. The other methods overlooked the fact and may lead to misleading results for competing causes.

Saddlepoint-Based Bootstrap Inference for Exponential Failure Times with Right-Censoring

✦Robert Paige, Missouri S & T, Rolla, MO 65409, *paigero@mst.edu*; Akim Adekpedjou, Missouri S & T; Noroharivelo Randrianampy, Missouri S & T

Key Words: Survival Analysis, Bootstrap, Saddlepoint

We develop a saddlepoint-based bootstrap method for making small-sample inference about the rate parameter of an exponential failure time in the presence of right-censoring for parametric or unspecified censoring time distributions.

344 Statistical methods for analyzing categorical data

Biometrics Section, Section on Health Policy Statistics, Section for Statistical Programmers and Analysts, Section on Statistical Consulting

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Determining Informative Cutpoints When Dichotomizing Data from the Northwestern University Spore in Prostate Cancer

✦Irene Helenowski, Northwestern University, , *i-helenowski@northwestern.edu*; Timothy Kuzel, Northwestern University; Borko Jovanovic, Northwestern University; Aishwarya Parameswaran , Northwestern University

Key Words: dichotomization

Maccallum et al (2002) discuss the importance of choosing cutpoints for dichotomizing continuous data cautiously. Here, we show the implications they make using PSA and BMI data. This work is funded by Grant P50 CA 090386.

Likelihood-Based Approach For Analysis Of Longitudinal Nominal Data Using Marginalized Random Effects Models

✦Keunbaik Lee, Louisiana State University Health Sciences Center, 1615 Poydras St., Suite 1400, New Orleans, LA 70112, *klee4@lsuhsc.edu*; Sanggil Kang, Sangji University; Xuefeng Liu, East Tennessee State University; Daekwan Seo, NCI/NIH

Key Words: marginal models, Quasi-Newton, Kronecker product

Likelihood-based marginalized models using random effects have become popular for analyzing longitudinal categorical data. These models permit direct interpretation of marginal mean parameters and characterize the serial dependence of longitudinal outcomes using random effects. In this paper, we propose model that expands the use of previous models to accommodate longitudinal nominal data. Random effects using a new covariance matrix with a Kronecker product composition are used to explain serial and categorical dependence. The Quasi-Newton algorithm is developed for estimation. These proposed methods are illustrated with a real data set and compared with other standard methods.

Double, Double: Exploiting Duality To Model Dispersion

✦Michael T Anderso, University of Texas at San Antonio, College of Business, Department of Management Science and Statistics, San Antonio, TX 78249, *michael.anderson@utsa.edu*

Key Words: COM-Poisson, duality, dispersion, GHPD, weighted distribution, Poisson regression

The COM-Poisson distribution of Conway and Maxwell is used in weighted Poisson regression to model phenomena which are significantly under- or over-dispersed. This flexibility in modeling dispersion derives from the property of duality inherent in the Poisson distribution, in distinction to most other discrete models. Other distributions also possess duality and provide for even more flexibility in modeling. Two will be discussed and examples of inference will be shown.

A Comparison Of Methods For Finding The Upper Confidence Limit For A Binomial Proportion When Zero Successes Are Observed

Courtney McCracken, Georgia Health Sciences University;
✦Stephen Looney, Georgia Health Sciences University, 1120 15th Street, AE - 1014, Dept. of Biostatistics, Augusta, GA 30912-4900, *slooney@georgiahealth.edu*

Key Words: Approximate methods, Binomial parameter, Exact methods, P-confidence

Confidence interval estimation for a binomial proportion is a long-debated topic, resulting in a wide range of exact and approximate methods. Many of these methods perform quite poorly when the number of observed successes in a sample of size n is zero. In this case, the main objective of the investigator is usually to obtain an upper bound, i.e., the upper limit of a one-sided confidence interval. Traditional notions of expected interval length and coverage probability are not applicable in this situation because it is assumed that the sample data have already been observed. In this paper we use p-confidence to evaluate nine methods for finding a confidence interval for a binomial proportion

when it is known that the number of observed successes is zero. We show that many popular approximate methods perform poorly based on p-confidence and recommend that the exact method be used.

A Linear Approximation Method For Parameter Estimation In Probit Regression

✦Haoyu Wang, Department of Statistics, University of California,Riverside, Riverside, CA 92521, *haoyu.wang@email.ucr.edu*; Subir Ghosh, Department of Statistics, University of California,Riverside

Key Words: Discrete Data, Linear Approximation, Maximum Likelihood, Probit Regression

In the Maximum Likelihood Estimation of the parameters in a probit regression model, we introduce a new approximate method to obtain the estimates. With our linear approximation, we find the exact solution of the Maximum Likelihood estimating equations. We compare our estimates with the standard numerical method estimates with a real data as well as a simulated data. We also present some theoretical properties of our estimates.

Classifying Time-Dependent Covariates In Modeling Correlated Data

✦Jeffrey Wilson, Arizona State University, , *jeffrey.wilson@asu.edu*; Anh Nguyen, Banner Good Samaritan Medical Center

Key Words: dependency, estimating equation, method of moments, longitudinal data

When analyzing longitudinal data it is essential to model both the correlation inherent from the repeated measures of the responses as well as the correlation created on account of the feedback created between the responses at a particular time and values of the predictors at other times. The generalized method of moment (GMM) for estimating the coefficients in longitudinal data with correct classification of time-dependent covariates provides substantial gains in efficiency over generalized estimating equations (GEE) with the independent working correlation. While the method provides advantages over generalized estimating equations with independent working correlation it requires correct classification into of the three types of time-dependent covariates. While most covariates in healthcare data may be type III it is determined if we misclassified them as type I or type II, as we saw in a subset of Arizona Medicare data 2003-05 on rehospitalization.

Existence Of Maximum Likelihood Estimates In Categorical Regression Models

✦Tezcan Ozrazgat Baslanti, University of Florida, Department of Statistics 102 Griffin-Floyd Hall, P.O. Box 118545, Gainesville, FL 32611, *tezcan@ufl.edu*; Michael Daniels, University of Florida; Alan Agresti, University of Florida

Key Words: existence of maximum likelihood estimates, categorical response data

It is important to know if the maximum likelihood estimates (MLE) exist or not for a given dataset as some software programs are unable to determine this. We review the existing literature on the existence of MLE for categorical response models such as the multinomial logistic model, the multinomial choice model, and cumulative link models of proportional odds form. We extend the definitions of separation and overlap, given by Albert and Anderson (1984), to the stereotype model and the adjacent-categories logit model. We summarize the existing literature, generalize some results, and examine connections between the different approaches. Combining ideas, we develop new methods of exploring if MLE exists.

345 Bayesian Modeling in Physics and Engineering

Section on Bayesian Statistical Science, Section on Quality and Productivity

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Bayesian Models for Image Object Classification with Application to Gold Nanoparticles

✦Bledar Konomi, Texas A&M University, 1501 Harvey Rd. ap. # 848, College Station, TX 77840, *alexandros@stat.tamu.edu*

Key Words: Object classification, Statistical shape analysis, Nanoparticles, Markov-chain Monte-carlo, RJ-MCMC, Monte Carlo Metropolis Hasting

By controlling the shape and size of nanoparticles during synthesis, one could control the properties of the synthesized material. Transmission electron microscopy imaging technique can be used to measure the morphological characteristics of nanoparticles, which can be simple circles or more complex irregular polygons with varying degree of scales and sizes. A major difficulty in analyzing the TEM images is the overlapping of objects having different morphological properties with no specific information about the number of objects present. Also the objects lying along the boundary render automated image analysis much more difficult. To overcome these challenges, we propose a Bayesian image segmentation and classification procedure based on the marked-point process representation of the objects. We derive models, both for the marks which parametrize the morphological aspects and the points which determine the location of the objects, to greatly reduce the complexity of the problem. We carry-out the inference by using a novel Markov chain Monte Carlo (MCMC) technique since the posterior distribution is analytically intractable.

Modeling And One-Sample Bayes Inference For The Wrapped-Maxwell-Boltzmann Uars Family Of Distributions On 3-D Rotations

✦Yu Qiu, Iowa State Univeristy, 50010, *yuqiu@iastate.edu*

Key Words: Circular Normal Distribution, Gibbs, Metropolis-Hasting, UARS, Wrapped Maxwell-Boltzmann, Wrapped Normal

We motivate a wrapped Maxwell-Boltzmann (wMB) ``spin" subclass of the UARS distributions on 3-D rotations studied by Bingham et al. (2009) as being a simple and good approximation of the limit distribution for the composition of many independent small rotations and having some advantages compared to other existing approximations. It is also analogous to the wrapped normal class of circular distribution in 2-D space. Non-informative Bayes methods are implemented via Markov Chain Monte Carlo methods and shown to have attractive frequentist properties. A particular interesting feature of the methods is

that depending upon the degree of wrapping involved, empirical convergence rates seem to vary from common $n^{-1/2}$ rates in small wrapping/ effectively regular cases to super-efficient n^{-1} rates in large wrapping/practically irregular cases.

Estimation Of Faraday Rotation Measures Of The Near Galactic Sky Using A Nonstationary Gaussian Process Model: Faster, Better Mcmc

✦Margaret Short, University of Alaska Fairbanks, P.O. Box 750125, Fairbanks, AK 99775, *mshort18@alaska.edu*; Philipp Kronberg, Los Alamos National Laboratory; Dave Higdon, Los Alamos National Laboratory

Key Words: spatial model, Faraday Rotation Measure, MCMC, process convolution

Our primary goal is to obtain a smoothed estimate of the magnetic field generated in and near to the Milky Way by using Faraday rotation measures (RMs). RMs provide an integrated measure of the effect of the magnetic field along the line of sight through the Milky Way to extragalactic radio sources. The ability to estimate the magnetic field generated locally by our Galaxy and its environs will help astronomers and astro-particle physicists distinguish local versus distant properties of the universe, and compute magnetic deflections of the (mysterious) ultra high energy [(> 10^19 eV)] cosmic rays after they enter the Milky Way. We model these data using Bayesian process convolution approach, fitted using MCMC. Our model incorporates nonstationarity by allowing model parameters to vary with galactic latitude. The MCMC is enhanced as well, to deal with the additional computational burden, due to the large data set and to the additional model parameters that are being estimated.

A Bayesian Approach To Detection Of Small Low Emission Sources

✦Xiaolei Xun, Texas A&M University, , *xxun@stat.tamu.edu*; Bani Mallick, Texas A & M University; Raymond James Carroll, Texas A&M University; Peter Kuchment, Texas A&M University

Key Words: Bayes factor, Model selection, Parallel tempering, Radiation source detection, Tomography

The article addresses the problem of detecting presence and location of a small low emission source inside of an object, when the background noise dominates. This problem arises, for instance, in some homeland security applications. The goal is to reach the signal-to-noise ratio (SNR) levels on the order of 0.001. A Bayesian approach to this problem is implemented in 2D. The method allows inference not only about the existence of the source, but also about its location. We derive Bayes factors for model selection and estimation of location based on Markov Chain Monte Carlo simulation. A simulation study shows that with sufficiently high total emission level, our method can effectively locate the source.

Bayesian Positioning Using Gaussian Mixture Models With Time-Varying Component Weights

✦Henri Pesonen, Tampere University of Technology, P.O.B. 553, Tampere, 33101 Finland, *henri.pesonen@tut.fi*; Robert PichÈ, Tampere University of Technology

Key Words: Bayesian filtering, Rao-Blackwellization, model uncertainty, multiple model filtering, positioning

Gaussian mixture models are often used in target tracking applications to take into account maneuvers in state dynamics or changing levels of observation noise. In this study it is assumed that the measurement or the state transition model can have two plausible candidates, as for example in positioning with line-of-sight or non-line-sight-signals. The plausibility described by the mixture component weight is modeled as a time-dependent random variable and is formulated as a Markov process with a heuristic model based on the Beta distribution. The proposed system can be used to approximate some well-known multiple model systems by tuning the parameter of the state transition distribution for the component weight. The posterior distribution of the state can be solved approximately using a Rao-Blackwellized particle filter. Simulations of GPS pedestrian tracking are used to test the proposed method. The results indicate that the new system is able to find the true models and its root mean square error-performance is comparable to filters that know the true models.

Emulators On Spatial Field Using Predictive Process

✦Anirban Mondal, Texas A&M University, Deapartment of Statistics, 3143 TAMU, College Station, TX 77843, *anirban@stat.tamu.edu*; Bani Mallick, Texas A & M University; Avishek Chakraborty, Texas A&M University; Yalchin Efendiev, Texas A&M University

Key Words: Bayesian approach to multivariate adaptive regression spline, predictive process, Bayesian inverse problems

The posterior distributions for Bayesian inverse problems are often intractable and we need thousands of MCMC samples from the posterior for the uncertainty analysis. The run times of the complex simulators are often such that it is prohibitively expensive to run those thousands of iterations in the MCMC procedure. One of the most commonly used approach to deal with this problem is based on building emulators. Kenedy and O'Hagen (2001), Oakley and O'Hagen (2002, 2004) used Gaussian process emulators for uncertainty and sensitivity analysis which can run almost instantaneously. The input parameter they considered is either known or a random variable (could be vector valued) which is calibrated using the emulator. We propose to extend the emulation based methods where the input is an unknown spatial field. We propose to use an emulator based on a Bayesian approach to multivariate adaptive regression spline (BMARS). By using predictive process an infinite dimensional spatial field is represented in terms of finite dimensional knot points and the covariance parameters and hence BMARS can be used where the knot points of the predictive process are used as regressors.

Parameter Estimation For Differential Equation Models: A Bayesian Approach

✦Kashyap Gupta, Texas A & M University, Department of Statistics,Texas A & M University, 3143 TAMU, College Station, TX 77843, *kashyap@stat.tamu.edu*

Key Words: Hierarchical Bayesian Model, Non Parametric Regression, Differential Equation, Inverse problem

Differential equations have a wide variety of applications in describing dynamic systems applied to scientific fields such as physics,engineering,economics and biomedical sciences. From a statistical point of view the inverse problem, using the measurements of state variables to estimate the parameters which characterize the system has not been explored much. Existing statistical methods which estimate parameters in differential equation models are frequentist in nature mainly dealing with plug in estimates of the state variables and its derivatives. In this paper we develop a hierarchical Bayesian model using traditional non-parametric regression techniques and a framework of measurement error. The proposed method and relevant results are applied using a flexible model called the FitzHugh-Nagumo model and an illustrative example has been presented.

346 Nonparametric Tests and Asymptotics ●

Section on Nonparametric Statistics, International Indian Statistical Association

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Power And Sample Size Computation With Binomial Proportions - Going Beyond Saw-Tooth Pattern

✦Anthiyur Kannappan, Cytel Inc., 675 Massachusetts Avenue, Cambridge, MA 02139, *kannappan@cytel.com*; Cyrus Mehta, Cytel Inc.; Pralay Senchaudhuri, Cytel Inc.

Key Words: power, sample, size, saw-tooth, binomial, proportion

It is well known that in the case of power & sample size computations with binomial proportions, there exists a phenomena of saw-tooth pattern in the power vs sample size curve. What this pattern indicates is that there is an 'anomalous' situation where the power of an experiment may decrease with an increased sample size. The reason for this seemingly anomalous situation is also known, that the discreteness of binomial distribution precludes computation of sample size to the exact value of alpha specified. Nevertheless, the overwhelming perception among statisticians is that saw-tooth pattern is an inherent anomaly present in power computation with binomial proportions. One way to counteract this perception is to present the computation of power not as a function of a single variable (viz.) sample size but as a function of two variables - sample size and 'attainable' alpha. This leads to a three dimensional graphical representation involving sample size, 'attainable' alpha, and power, where it can be easily visualized and understood that there is no anomalous situation present in power computations with binomial proportions. This paper presents several illustrative examples.

Intrinsic Regression Models For Lie Group-Valued Data

✦Emil A. Cornea, University of North Carolina, Dept. of Biostatistics, School of Public Health, Dept. of Biostatistics, CB# 7420, University of North Carolina, Chapel Hill, NC 27599-7420, *ecornea@bios.unc.edu*; Hongtu Zhu, University of North Carolina Department of Biostatistics; Joseph G. Ibrahim, University of North Carolina

Key Words: semiparametric regression model, Lie group, link function, conditional moment, medical imaging

In many fields, like medical imaging, computational biology, computer vision, there is a growing interest in analyzing various data, such as 3D rotations, symmetric positive-definite matrices, and diffusion tensors, and their association with a set of Euclidean covariates. These data do not form a vector space and may rather be viewed as lying in appropriate Lie groups, or more generally, Riemannian manifolds, thus directly applying classical multivariate regression may be inadequate. The aim of this work is to present an intrinsic regression model for the analysis Lie group-valued data and the association with a set of covariates in a Euclidean space. Our model is semiparametric, solely based on the conditional first-order moment, avoiding specifying any parametric distributions; it uses a link function to map covariates from the Euclidean space to the Lie group of responses. We develop a two-stage procedure to calculate the parameter estimates, and determine their asymptotic distributions. We construct goodness-of-fit statistics for testing possible model misspecifications, and score statistics to test hypotheses on unknown parameters. Applications in neuroimaging are presented.

Semiparametric Function Estimation Using Shrinkage Techniques

✦Mohamed Amezziane, DePaul University, 60660, *mamezzia@depaul.edu*; Syed Ejaz Ahmed, University of Windsor

Key Words: pre-test estimators, nonparametric estimation, shrinkage estimation, smoothing parameter, shrinkage coefficient

We use shrinkage techniques to develop a class of semiparametric estimators of functions (distribution, density, regression, etc.) which can be obtained as linear combinations of fully determined parametric functions and nonparametric function estimators. We present the asymptotic properties of the proposed class of estimators and compare their performance to that of classical nonparametric estimators. Moreover, we show that the proposed estimators do not require the use of optimally selected smoothing parameters and are therefore less sensitive to the effect of curse of dimensionality.

A Nonparametric Test Of Missing Completely At Random For Incomplete Multivariate Data

✦Yao Yu, University of California, Riverside, 92507, *yao.yu@email.ucr.edu*; Jun Li, University of California, Riverside

Key Words: missing data, k-sample test, nonparametric test, multivariate imputation

Missing data occur in many real world studies. Knowing the type of missing mechanisms is critical for adopting the appropriate statistical procedure. Many statistical procedures assume missing completely at random (MCAR) due to its simplicity. Therefore, it is extremely important to test whether this assumption is satisfied before applying those procedures. In the literature, most of the procedures for testing MCAR were developed under normality assumption, which is sometimes difficult to justify in practice. A nonparametric procedure which does not require distributional assumptions is more desirable. In this talk, we propose a nonparametric test of MCAR for incomplete multivariate data. The proposed test first employs appropriate missing data imputation method to impute missing data, and then applies the nonparamet-

ric multiple sample homogeneity test to the imputed data. The performance of the proposed procedure is evaluated by a simulation study, which shows that our procedure has the Type I error well controlled at the nominal level and also has reasonable power against a variety of alternatives.

Testing For The Covariate Effect In The Fully Nonparametric Ancova Model

✦SHU-MIN LIAO, Amherst College, Amherst Colleg Box 2239, P.O. 5000, Amherst, MA 01002-5000, *sliao@amherst.edu*; Michael G. Akritas, Pennsylvania State University

Key Words: Nonparametric, Analysis of Covariance, Nested designs, Asymptotic theory

In this talk, we introduce a new approach for testing the covariate effect in the context of the fully nonparametric ANCOVA model which capitalizes on the connection to the testing problems in nested designs. The basic idea behind the proposed method is to think of each distinct covariate value as a level of a sub-class nested in each group/class. A projection-based tool is developed to obtain a new class of quadratic forms, whose asymptotic behavior is then studied to establish the limiting distributions of the proposed test statistic under the null hypothesis and local alternatives. Simulation studies show that this new method, compared with existing alternatives, has better power properties and achieves the nominal level under violations of the classical assumptions. Analysis of three real data sets are also included.

The Efficiency Of The Second-Order Nonlinear Least Squares Estimator And Its Extension

✦Mi Jeong Kim, Texas A&M University, Department of Statistics, Texas A&M University, College Station, TX 77843, *mjkim@stat.tamu.edu*

Key Words: Second-order least squares estimator, heteroscedasticity, moments, semiparametric methods

We revisit the second-order nonlinear least square estimator proposed in Wang and Leblanc (2008) and show that the estimator reaches the asymptotic optimality concerning the estimation variability. Using a fully semiparametric approach, we further modify and extend the method to the heteroscedastic error models and propose a semiparametric efficient estimator in this more general setting. Numerical results are provided to support the results and illustrate the finite sample performance of the proposed estimator.

Asymptotic Efficiency Of Ridge Estimator In Linear And Semiparametric Linear Models

✦June Luo, Clemson University, 229 Barre hall, Clemson, SC 29631, *jluo@clemson.edu*

Key Words: high dimension, ridge regression, differencing sequence

The linear model with a growing number of predictors arises in many contemporary scientific endeavor. In this article, we consider the commonly used ridge estimator in linear models. We propose analyzing the ridge estimator for a finite sample size n and a growing dimension p. The existence and asymptotic normality of the ridge estimator are established under some regularity conditions when p goes to infinity. It also occurs that a strictly linear model is inadequate when some of the relations are believed to be of certain linear form while others are not easily parameterized, and thus a semiparametric partial linear model is considered. For these semiparametric partial linear models with p > n, we develop a procedure to estimate the linear coefficients as if the nonparametric part is not present. The asymptotic efficiency of the proposed estimator for the linear component is studied for p going to infinity.

347 Classification and Clustering ■

Section on Statistical Computing, International Indian Statistical Association, SSC

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Can The Carnegie Research University Classification Methodology Be Improved?

✦David W. Scott, Rice University, , *scottdw@rice.edu*

Key Words: University rankings, principcal components, classification

In January, 2011, the Carnegie Foundation issued its 2010 university classifications, which were last updated in 2005. Among the various categories, some 500 research universities were classified as either (1) RU/VH: Research Universities / very high research activity; (2) RU/H: Research Universities / high research activity; or (3) DRU: Doctoral / Research Universities. The methodology used by the Carnegie Foundation strives to strike a balance between raw size and quality of research and institutions by constructing two separate principal components: one based upon aggregate data and the other using per capita data. In this talk, we consider the choices made and investigate alternatives that are available. Recommendations are evaluated.

On The Effect And Remedies Of Shrinkage On Classification

✦Chong Zhang, Department of Statistics and Operation Research, UNC-Chapel Hill, CB# 3260, UNC, Chapel Hill, NC 27599, *chongz@email.unc.edu*; Yufeng Liu, Department of Statistics and Operation Research, UNC-Chapel Hill

Key Words: classification probability estimation, penalized large margin classifier

Shrinkage methods have been shown to be effective for classification problems. As a form of regularization, shrinkage through penalization helps to avoid overfitting and produce accurate classifiers for prediction, especially when the dimension is relatively high. Despite the benefit of shrinkage on classification accuracy of resulting classifiers, in this paper, we demonstrate that shrinkage creates biases on classification probability estimation. In many cases, this bias can be large and consequently yield poor class probability estimation when the sample size is small or moderate. We offer some theoretical insight on the effect of shrinkage and provide remedies for better class probability estimation. Using penalized logistic regression, we demonstrate that our proposed refit method gives similar classification accuracy yet remarkable improvements on probability estimation on several simulated and real data examples.

Supervised Clustering Methods For High Dimensional Data

✦Maria Reynolds, University of North Carolina at Chapel Hill, 130 Marlowe Court, Carrboro, NC 27510, *mariaereynolds@gmail.com*; Eric Bair, University of North Carolina-Chapel Hill

Key Words: clustering, high dimensional, microarray, semi supervised, genetic, machine learning

Traditional clustering techniques are purely unsupervised and do not consider an outcome variable. In many situations, it would be useful to incorporate information from an outcome variable when performing clustering. For example, in microarray studies, hierarchical clustering is often used to partition tumors into subtypes, and it is hoped that patients with different subtypes will have different prognoses. Similarly, in case/control genetic studies, some of the "controls" may later develop the disease. It would be desirable to identify controls that have similar genetics to the cases and hence are more likely to become cases in the future. When clustering high dimensional data sets, transforming variables and/or applying feature selection can often improve the results. Several methods have been proposed for feature selection when performing clustering, but these methods are unable to account for an outcome variable. We demonstrate that semi-supervised methods, which can account for an outcome variable, improve on existing feature selection methods when applied to both simulated and real data sets.

Semi-Supervised Methods For Classification In The Presence Of Mislabeled Outcome Variables

✦Rebecca Susan Rothwell, University of North Carolina Chapel Hill, 115 Stinson Street, Chapel Hill, NC 27514, *rrothwel@email.unc.edu*; Eric Bair, University of North Carolina-Chapel Hill

Key Words: classification, machine learning, high-dimensional, microarry, semi-supervised, genetics

Mislabeled outcome data is very common in classification problems. For example, in case-control genetic studies, frequently some of the "controls" will become cases in the future. Likewise, many cancer studies use DNA microarrays to try to identify subtypes of cancer with different prognoses. However, these cancer subtypes cannot be observed directly, so surrogate variables (such as survival times) are often used as the outcome of interest, and it is unlikely that these survival times are perfectly correlated with the unobserved cancer subtypes. Therefore it is important to have classification methods that are robust against mislabeled data. Most existing classification methods do not meet this requirement. We show that semi-supervised methods, including supervised principal components, are more robust in these situations. We apply our methods to a series of simulated and real classification problems involving high-dimensional data and demonstrate that our methods produce a significantly lower misclassification error rate than conventional classification procedures.

Three-Way Nonmetric Multidimensional Scaling For Fusing Heterogeneous Data Sources

✦Brent S Castle, Indiana University, 5347 N. College Ave. #315, Indianapolis, IN 46220, *bscastle@cs.indiana.edu*; Michael Trosset, Indiana University

Key Words: three-way multidimensional scaling, nonmetric multidimensional scaling, data fusion, multi-view learning, raw stress criterion, partial ordering

We consider the problem of classification given several heterogeneous sources of data. For example, images on the popular image hosting website Flickr are often accompanied by captions, comments, and tags. One might be able to improve classification performance by using the images as well as the various text sources. A popular approach to classification in this context is to form classifiers using pairwise comparisons from each of the data types and to combine the resulting predictions. In this work we propose a three-way nonmetric multidimensional scaling formulation as a means to combine proximity measures constructed from each data type. We first embed the objects into a single low-dimensional Euclidean representation and subsequently perform classification using the resulting configuration. This nonmetric multidimensional scaling algorithm uses only ordinal properties of each of the proximity measures so one can remain agnostic to the measure itself while adhering to its ordinal properties.

Bayesian Factor Analysis For Clustered Categorical Data

✦Taiyeong Lee, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513, *taiyeong.lee@sas.com*; Yongdai Kim, Seoul National University

Key Words: Bayesian Factor Analysis, MCMC algorithm, Market Basket Analysis, Factor Model

We propose a Bayesian factor model for clustered binary data, which can be used for market basket analysis where each cluster corresponds to each customer, and binary vectors in each cluster represent the shopping history of the corresponding customer. We use latent variables for modeling dependency of binary vectors. That is, a vector of binary random variables is obtained by making the threshold of a vector of latent variables, which are assumed to be Gaussian random variables, at 0. Then we construct a factor model of latent variables. The proposed model is characterized by that each cluster has its own factor model, but the parameters can be shared across the clusters in the model. An efficient MCMC algorithm is developed and the method is illustrated on a real data set of market basket analysis

348 Computer Experiments and Statistical Forecasting ■●

Section on Physical and Engineering Sciences, Section on Quality and Productivity

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Adaptive Probability-Based Latin Hypercube Designs

✦Ying Hung, Rutgers, the state university of New Jersey, 110 Frelinghuysen Road, Piscataway, NJ 08854, *yhung@stat.rutgers.edu*

Key Words: Adaptive cluster sampling, Computer experiment, Latin hypercube design, Sequential design, Space-filling design

Adaptive sampling is an effective method developed mainly for regular regions. However, experimental regions in irregular shapes are commonly observed in practice. Motivated by a data center thermal management study, a new class of adaptive designs is proposed to accommodate a specific type of irregular regions. Because the adaptive procedure introduces biases into conventional estimators, several design-unbiased estimators are given for estimating the population mean. Efficient and easy-to-compute unbiased estimators are also introduced. The proposed method is applied to obtain an adaptive sensor placement plan to monitor and study the thermal distribution in a data center.

Choice Of Initial Sample Size For The Sequential Analysis Of A Computer Experiment

✦Dexter C. Whittinghill, Rowan University, Department of Mathematics, 201 Mullica Hill Rd., Glassboro, NJ 08028, *whittinghill@rowan.edu*

Key Words: computer experiments, sequential analysis, initial sample size, maximin Latin hypercube, expected improvement, cross-validation

Computer experiments, which comprise the evaluation of large, complicated mathematical models using computer code, continue to become more and more prominent in applied research as computers improve, and the body of literature on the design and analysis of computer experiments expands. While early research investigated the design and analysis for fixed sample sizes for the number of runs of the computer experiment, lately more and more investigators look at sequential experiments for optimization (Lam & Notz 2008, Williams, Santner & Notz 2000), determination of contours (Ranjan, Bingham & Michailidis 2008, Roy & Notz 2010) and calibration (Kumar & Notz 2010). However, guidance for what fraction of the budgeted sample size should be used for the initial sample size is largely unexplored. We will investigate the effect that the choice of different initial sample sizes has on the accuracy in optimization and the determination of contours.

Genuine Exact Two-Stage Methodologies For Producing Assigned Accuracy Estimators For A Gamma Mean

✦Kevin Paul Tolliver, United States Census Bureau, , *kptolliver@gmail.com*

Key Words: Gamma, sequential estimation, two-stage, mean time to failure, genuine, risk

This paper proposes two methods for finding estimators with assigned accuracy: (1) point estimator and (2) an interval estimator. It implements a genuine two-stage sampling procedure. The term genuine refers to the fact, that in contrast to previous methods, the procedures proposed herein are based on the combined samples from both stages, rather than ignoring the data from the first-stage sample. In addition all the results are exact. At no point was an asymptotic or large sampling approximation used and all the derivations assumed an underlying distribution of Gamma. Results are found for when shape is both known and unknown.

Off-Line Data Analysis And Real-Time Fault Detection And Isolation With Distributed Processing And State Space Compression

✦Spencer Graves, Structure Inspection and Monitoring, Inc., 751 Emerson Ct., San Jose, CA 95126, *spencer.graves@structuremonitoring.com*

Key Words: Data compression, structural health monitoring, Bayesian sequential updating, Distributed filtering, critical infrastructure, environmental protection

The quantity of data that could be collected in any real-time monitoring application is limited by the costs of data collection, transmission and storage. Smart sensors allow computations to be done anywhere, and analog signals can theoretically be digitized to any number of bits with any sampling frequency. The information obtained for a given budget can be maximized by considering what features of the data are most informative, and where those features can be most cheaply encoded. The high order bits in a digital representation almost never change, and the lowest order bits may be noise. Moreover, only change is informative. To determine the number of bits to retain, we consider "observation = important + unimportant + noise". We model this decomposition with Bayesian state space structures, reporting a new state only when predictions from the last report are not adequate. To support off-line analysis for model improvement, we also report outliers and samples of other observations. This creates a need to develop new methods to analyze data in this compressed format.

Parallel Computations in R, with Applications for Statistical Forecasting

Murray Stokely, Google; ✦Farzan Rohani, Google, , *farzan@google.com*; Nate Coehlo, Google; Eric Tassone, Google

Key Words: R, distributed computing, forecasting, simulation, statistical computing

We demonstrate the utility of massively parallel computational infrastructure for statistical computing with an implementation of the MapReduce paradigm for R. This allows users to write computations in a high-level language that are then broken up and distributed to worker tasks in Google datacenters. Results are collected in a scalable, distributed data store and returned to the interactive user session. We present real-world results of failures, run times, and performance from multiple applications to allow practitioners to better understand the nature of massively parallel statistical simulations. We motivate this with a forecasting application that fits a variety of models, prohibiting an analytical description of the statistical uncertainty associated with the overall forecast. To overcome this, we generate simulation-based uncertainty bands, which necessitates a large number of computationally intensive realizations. Our technique cut total run time by a factor of 300. This permits analysts to focus on statistical issues while answering questions that would be intractable without significant parallel computational infrastructure.

Evaluating The Robustness Of The Regression Models Obtained In Environmental Sciences

✦Fred J. Rispoli, Dowling College, 150 Idle Hour Blvd, Oakdale, NY 11769, *rispolif@dowling.edu*

Key Words: design of experiments, multiple regression

In environmental sciences, it is a standard practice to use statistical design of experiment (DOE) to study the influence of multiple parameters (physical, chemical or biological) on the toxicity of a contaminant or to evaluate the efficiency of an environmental process under a wide range of conditions. The objective of this study is to investigate the robustness of regression models based on the computer aided design and analysis of experiments often used in environmental biotechnology. We are particularly interested in how well the design and analysis process works when we randomly introduce a small "error" into the design points as a way of representing measurement error. The perturbed designs are then used together with associated mathematical models obtained from the original designs to simulate experiments. Regression analysis is used to produce a new model which is compared to the original model. The process allows us to study the robustness of experimental designs as well as vary experimental parameters to determine conditions that will improve coefficient reliability. Thus make more reliable inferences.

Calibrated Probabilistic Forecasting Of Extreme Events Based On Physical Model Outputs

✦Hongfei Li, IBM Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY 10598, *liho@us.ibm.com*; Huijing Jiang, IBM Thomas J. Watson Research Center; Jonathan R. M. Hosking, IBM T. J. Watson Research Center

Key Words: Extreme events, Physical model outputs, Heavy tail, gust speed

In certain circumstances, physical model outputs cannot provide direct information used in real applications. For example, power outages caused by wind gusts are major concerns for electric utility companies, but some meteorological models produce forecasts only of sustained wind speed instead of instantaneous gust. We propose a Bayesian hierarchical model to forecast the gust speed utilizing wind speed forecasts given by multiple meteorological model outputs, and apply it to real data. The weather forecast outputs are generated by IBM Deep Thunder system which utilizes WRF-ARW to enable effective forecasts with up to 72 hours lead time, on 28x17 grids with 2 km resolution across the area of Westchester county, NY. Particular attention is given to characterizing the heavy tailed properties of gust speed using Rayleigh distribution. The parameters of Rayleigh distributions are modeled through multiple wind model outputs and calibrated using gust observations collected at limited weather stations.

349 Survey Participation: Maximizing Cooperation and Minimizing Burden

Section on Survey Research Methods, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Cell Phone Versus Landline Respondents: Who'S More Cooperative After Screening?

✦Alicia M. Frasier, NORC at the University of Chicago, 55 E. Monroe, Chicago, IL 60603, *frasier-alicia@norc.org*; Heather M. Morrison, NORC at the University of Chicago

Key Words: RDD, Eligibility Screening, Cell Telephones

As landline telephone coverage declines nationally, surveys are supplementing landline random-digit-dial (RDD) samples with cell phone RDD samples to minimize potential bias from non-coverage of wireless-only households. This may increase respondent burden (through added screening questions designed to ensure respondent safety and confirm cell phone eligibility) and decrease participation rates. While initially gaining respondent cooperation may prove challenging, there is evidence to indicate that those cell respondents who do complete eligibility screening are more cooperative than their landline counterparts. By examining data from the cell component of a large population-based health survey we see that once cooperation is gained and initial screening completed, cell phone respondents require less time to complete the survey - potentially indicating they are more cooperative than landline respondents once trust is gained. We compare interview duration of cell phone and landline respondents, looking at timings before and after eligibility determination, to ascertain if the shorter interview duration observed in cell phone interviews may be attributed to cooperation propensity.

Given Different Mobile Phone Device Data Collection Options - What Will Respondents Choose?

✦Daniel Evan Williams, Opinionology, 701 East Timpanogos Parkway, Building M, Orem, UT 84097, *dwilliams@opinionology.com*; Laura Haber, Opinionology

Key Words: Mobile phone data collection, At event sampling, Mode effects, Text Message surveys, Mobile web surveys, IVR data collection

Over a five day period at a major sporting event, spectators were invited to use their mobile phone to respond to a brief survey on their experience. Options for mobile web, interactive voice recording (IVR) and SMS text messaging were given so a respondent could choose which way to participate in the survey. Participation is tracked and data differences by questions are examined to determine mode affects. Applications are relevant for at event data collection and potential data differences that may exist from the different modes.

Development And Validation Of A Questionnaire To Measure Respondents' Perceived Risks Of Web Surveys

✦Hsien-Yuan Hsu, National Taiwan Normal University, No.129, Sec. 1, Heping E. Rd., Da'an Dist., Taipei, International 106 Taiwan, *hsuhy@ntnu.edu.tw*; Yi-Hua Lai, National Taiwan Normal University; Hsin-Ying Chin, National Taiwan Normal University

Key Words: measurement development, perceived risk theory, response rate, web surveys

Based on Perceived Risk Theory, we hypothesized that privacy and time risks of Web surveys as perceived by respondents would have negative impacts on response rates. However, no tool for measuring respondents' perceived risks of Web surveys was found. The purpose of this study was to develop a questionnaire which could measure respondents' perceived privacy and time risks of Web surveys. Privacy risk was defined as "potential loss of control over personal information or the responses of Web surveys", while time risk was defined as " potential loss of time when replying a Web survey or spending more time on follow-up Web surveys after finishing the current one". The Questionnaire of

Web Survey Perceived Risk (QWSPR) was developed and comprised 10 items. Survey data consisting of 620 college students' responses to QWSPR were randomly separated into two groups for cross validation of score validity: group one was used for exploratory factor analysis and group two was used for confirmatory factor analysis. Due to cross-loading problems, one item was deleted. Results indicated that a two-factor (privacy and time risks) structural model was validated.

Testing Gender Difference In Perceived Time Risk Of The Web Surveys Using Mimic Model

✦Hsin-Ying Chin, National Taiwan Normal University, No.129, Sec. 1, Heping E. Rd., Da'an Dist., Taipei, International 106 Taiwan, *ykingamy@yahoo.com.tw*; Hsien-Yuan Hsu, National Taiwan Normal University; Yi-Hua Lai, National Taiwan Normal University

Key Words: gender difference, perceived risk theory, response rate, web surveys

Perceived Risk Theory (PRT) has been applied to explain Web survey respondents' reply intention. Prior research found that respondents who perceived higher level of time risk of Web survey (i.e., potential loss of time when replying a Web survey or spending more time on follow-up Web surveys after finishing the current one) were less likely to reply to academic Web surveys. No study to date has investigated why respondents perceived different levels of time risk. The purpose of this study was to examine whether gender differentiates perceived the time risk of Web surveys. A subscale (4 items) of Questionnaire of Web Survey Perceived Risk (QWSPR) was utilized to measure 620 college students' time risk perception in a survey deployed in fall 2010. A multiple-indicator multiple-cause (MIMIC) model was used to assess the difference between male and female students' perceived time risk. Results indicated that female students perceived a higher level of time risk than male students. Our findings raised the issue of how to develop an effective strategy which could decrease female respondents' time risk perception in order to enhance the Web survey response rate.

The Impacts Of Respondents' Perceived Privacy And Time Risks On The Web Surveys Reply Intention

✦Yi-Hua Lai, National Taiwan Normal University, No.129, Sec. 1, Heping E. Rd., Da'an Dist., Taipei, International 106 Taipei, *evalai920@gmail.com*; Hsin-Ying Chin, National Taiwan Normal University; Hsien-Yuan Hsu, National Taiwan Normal University

Key Words: perceived risk theory, response rate, web surveys

Perceived Risk Theory (PRT) has been widely applied to predict consumers' online purchase intention. This study employed PRT to explain low response rates confronting by Web surveys. The Questionnaire of Web Survey Perceived Risk (QWSPR) comprised of 9 items was utilized to measure 620 college students' perceived privacy risk and time risk. Privacy risk was defined as "potential loss of control over personal information or the responses of Web surveys" (measured by 4 items), while time risk was defined as "potential loss of time when replying a Web survey or spending more time on follow-up Web surveys after finishing the current one" (measured by 5 items). In addition, students were asked to evaluate their reply intention if they received an academic Web survey via email. Student characteristics were also collected. After controlling for students' gender, age, parents' education, family income, and major, results showed as time risk perceived by students increased, Web survey reply intention decreased. However, no effect was found for perceived privacy risk on reply intention. Future study is needed to investigate how to decrease respondents' perceived time risk of Web surveys.

Managing Response Burden By Controlling Sample Selection And Survey Coverage

✦Sebastien Landry, Statistics Canada, 100 Tunney's Pasture Driveway, R.H.Coats Building, 11-P, Ottawa, ON K1A 0T6 Canada, *sebastien.landry@statcan.gc.ca*

Key Words: Business surveys, Response burden, Sample selection, Take-none strata

Statistical agencies are constantly making efforts to control the response burden of their household and business survey respondents. Statistics Canada's Survey on Employment, Payroll and Hours is no exception. This monthly business survey, which produces estimates and determines the month-to-month changes for variables such as employment, earnings and hours at detailed industrial levels for Canada, the provinces and the territories, currently manages response burden by making use of administrative data and by having rules that prevent establishments from rotating in the sample too soon after being rotated out. Recently, two new ideas to decrease even more the response burden for respondents to this survey have been studied. The first is to control the overlap of the samples from one month to the next by the use of the microstrata method (RiviËre (2001)) in the sample selection process. The second is the increased number of establishments in the take-none strata. This paper will present the studies that evaluated the pros and cons of implementing each of these new features in the survey.

Beyond Response Rates: Exploring the Impact of Prepaid Cash Incentives on Multiple Indicators of Data Quality

✦Rebecca Medway, Joint Program in Survey Methodology, University of Maryland, , *rmedway@survey.umd.edu*; Roger Tourangeau, Joint Program in Survey Methodology

Key Words: incentives, data quality

As survey response rates continue to decline, incentives are widely used as a way to motivate sample members to respond. However, rather than simply being satisfied with increased response rates, it is also important to determine whether the use of incentives affects the quality of the resulting survey data. Several indicators of data quality that are widely used in survey research have not often been analyzed in the context of incentive experiments. For example, little is known about the effect that incentives have on the prevalence of satisficing behaviors, such as nondifferentiation or response order effects. Furthermore, the effect of incentives on the reliability of survey responses is not well understood. Using data from an incentive experiment included in a recent mixed-mode survey, this presentation will explore the effect of prepaid cash incentives on these, and other, indicators of data quality. Differential effects of incentives on data quality across modes and demographic subgroups will also be discussed.

350 Adaptive Bayesian Trials ■

Biopharmaceutical Section, Section on Bayesian Statistical Science
Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Beyond Bells And Whistles In Bayesian Sample Size Computation

✦Mani Lakshminarayanan, Merck & Co, , *Mani_Lakshminarayanan@merck.com*; Xin Zhao, Merck & Co

Key Words: Bayesian Power, Decision Analysis, Predictive Probability, Conditional Power

Computing sample size is one of the key efforts that need to take place at the design stage of any statistical experiment including clinical trial development. Suppose our primary interest is to compare two treatments based on a difference (D) in a pre-specified endpoint. Frequentist or classical statistical approach to sample size determination starts with finding the number of patients required with a prespecified power (for a fixed Type I error rate) to detect the difference D. All these assumptions that need to be made before estimating a sample size have contributed to the general agreement that a formal sample size computation in clinical trials (or other statistical experiments) is very subjective. A Bayesian approach provides a natural framework for incorporating any subjective information including the prior distributions for all the parameters. A Bayesian sample size computation provides multitude of options that are easily adaptable to underlying formulation of the problem. In this presentation, we will properties of approaches for both Bayesian sample size computation and reestimation and contrast them with frequentist alternatives.

Bayesian Adaptively Randomized Clinical Trial Of Late Stage Non-Small Cell Lung Cancer

✦Chunyan Cai, University of Texas MD Anderson Cancer Center, ivision of Quantitative Sciences - Unit 1409, P. O. Box 301402, Houston, TX 77230-1402, *cyretni@gmail.com*; Valen E. Johnson, University of Texas MD Anderson Cancer Center; Ying Yuan, University of Texas MD Anderson Cancer Center

Key Words: Bayesian adaptive randomization, Bayes factor, Symptom interventions, Factorial design

Bayesian adaptively randomized clinical trials assign patients to treatments with probabilities that are calculated using the outcomes of previous patients. Recently, the use of Bayesian adaptive trials has increased due to their potential for increasing the number of patients assigned to efficacious treatments. In this proposal, we develop a Bayesian adaptive randomization design to test the efficacy of 16 combinations of 4 trial agents for reducing symptoms of late stage non-small cell lung cancer (NSCLC) patients. To obtain initial estimates of treatments effects, we assign the first 32 patients to treatments following a randomized factorial design; subsequent patients are assigned to treatments according to the posterior probability that each treatment is most efficacious. Bayes factors are used in the computation of posterior probabilities in the test-based design. Compared with equal randomization design, we show that our Bayesian adaptively randomized design assigns more patients to better treatments and efficiently identify the best treatment from multiple treatments.

Dose-Finding For Combination Trials Based On Toxicity Probability Intervals

✦Fang Liu, Octagon Research Solution, Inc., 585 East Swedesford Road, Wayne, PA 19087, *fliu@octagonresearch.com*; ✦Zijiang Yang, TechData Service Company, LLC, 700 American Ave, Ste102, King of Prussia, PA 19406, *zjyang@temple.edu*

Key Words: Adaptive design, Dose-finding, Bayesian design, Combination trials, Dose escalation

When two drugs are combined in the drug development, the toxicity rate for the combo is expected to be higher than the individual drug and there could be multiple combination levels yielding the same toxicity rate, i.e. the MTD is not unique. It is crucial to identify the contour lines with the same toxicity rate and carry the most efficacious combo to later phase trials. This paper proposed a new dose-finding design for combination trials, which extended the design proposed by Ji's simple Bayesian design for mono-therapy. The new design is based on a beta/binomial Bayesian model and uses a zone-based escalation rule which can extend to all pre-specified dose combinations. The new design can be easily implemented by physicians without any input from statistician during the escalation process. Simulation studies indicate that the new design outperforms the traditional combo escalation design and competitive with the complicated CRM-based designs.

Application Of Bayesian Predictive Probability Approach In A Two-Stage Clinical Proof Of Concept Trial

✦Feng Gao, GlaxoSmithKline, Wayne, PA 19087, *feng.f.gao@gsk.com*

Key Words: adaptive, Bayesian, POC, predictive, decision making

The Bayesian predictive probability approach has been used in various trials in recent years, especially in early phase trials. The predictive idea fits the exploratory nature of early phase trials very well and the Bayesian predictive procedure provides an attractive tool to evaluate the chance if a trial will end up with a positive result or not. In this paper, an application of Bayesian predictive probability approach to an actual two-stage clinical Proof of Concept trial will be presented in terms of design of the trial, monitoring the trial and go/no go decision making. This model-based, Bayesian adaptive design provided the study team a tool to maximize the information that can be obtained from the trial and helped the study team's decision while spending money utilizing resource more efficiently.

Bayesian Adaptive Randomization Design Using Posterior Predictive Probability

✦Xuemin Gu, Eli Lilly and Company, , *xuemin_gu@lilly.com*; Brenda Gaydos, Eli Lilly

Key Words: Adaptive Design, Adaptive Randomization, Predictive Probability

Outcome-based adaptive randomization has been increasingly used as an adaptation element in Bayesian adaptive design for clinical trials, especially in the exploratory phase of drug development, where the most common randomization schemes involve using patient allocation ratios proportional to the probability of one treatment being better than all the others or posterior point estimate of treatment efficacy after proper

scaling. However during the planning stage of a clinical trial, these methods, intuitively appealing, all show great variations of allocation ratios in the early stage of the trial. Various remedies for the early variation are ad hoc. In the present study, we show that the use of predictive probability approach can solve the problem satisfactorily. Additionally, other benefits of predictive probability approach are demonstrated. In particular, for example one set of decision rule will be used for both early stopping and final decision, which makes the decision process more consistent.

Bayesian Response Adaptive Randomization For Delayed Response With A Short-Term Outcome

✦Mi-Ok Kim, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave., MLC 5041, Cincinnati, OH 45229 U.S., *MiOk.Kim@cchmc.org*; Chunyan Liu, Cincinnati Children's Hospital Medical Center; J. Jack Lee, Univ of Texas - MD Anderson Cancer Center

Key Words: Bayesian, Response Adaptive Design, clinical trial

We consider an application of response-adaptive randomization (RAR) design in a clinical trial where the primary endpoint takes a long time to observe but a short-term "surrogate" outcome is available. The asymptotic properties of the design have been shown to be little affected when the delay is not very long relative to the accrual process (e.g. Bai et al, 2003; Hu et al, 2008). These theoretical results, however, are not useful when the delay is long, as in many survival outcome trials, or the sample size is small. Huang et al (2009) proposed a Bayesian approach of utilizing the short-term outcome for a RAR for the long-term response. We use Huang et al's approach to study the effect of the delay on a RAR design with a small n and to investigate when Huang et al (2009)'s approach is beneficial compared to an approach that does not utilize the short-term outcome. Simulation results show that the more complex Huang et al's approach performs better as the priors becomes more informative and/or if the treat effect differs by the short-term outcome.

A Comparison Of Covariate-Adaptive Randomization Procedures In Clinical Trials

Dan Neal, University of Florida Department of Biostatistics; Tamekia Jones, University of Florida Department of Biostatistics; Jivan Ginosian, Bioness Inc.; ✦Samuel S. Wu, University of Florida Department of Biostatistics, 1329 SW 16th Street Room 5231, Gainesville, FL 32601, *samwu@biostat.ufl.edu*

Key Words: adaptive-covariate randomization, clinical trial design, Wei's urn, Pocock-Simon

When there are many important covariates or prognostic factors in a clinical trial, stratified randomization becomes infeasible. To ensure balance over the covariates, several authors have proposed covariate-adaptive randomization procedures, which are especially appealing when there are pre-planned subgroup analyses involving the covariates. In this talk, we will compare Zelen's rule, the Pocock-Simon procedure, Wei's marginal urn design and other methods on their balancing properties and predictability of assignment. The results will be illustrated using a real randomized controlled trial that was conducted to determine the effectiveness of electrical stimulation therapy in chronic post-stroke subjects.

351 Issues in Early Development and Pharmacokentic Studies ■

Biopharmaceutical Section

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

When Is A Biomarker-Based Study Design In Drug Development Likely To Succeed?

✦Deepak B. Khatry, MedImmune, One Medimmune Way, Gaithersburg, MD 20878, *KhatryD@Medimmune.com*

Key Words: Biomarker, Diagnostics, Adaptive design, Clinical trial, Personalized healthcare, Stratified medicine

Advances in technology enable us to measure previously unquantifiable biological products with increasing precision, promising an era of personalized healthcare. Biomarker tools, if properly utilized as diagnostic (Dx) devices, can increase the overall efficiency in bringing a new therapy to market. Such an increase in efficiency can benefit drug industry and patients alike by reducing development cost, increasing probability of trial success, and increasing benefit/risk ratio for targeted patient subgroups. An important question is when to pursue a biomarker-based design over the traditional randomized control trial (RCT) design. Aside from considerations of market share and profits, there are statistical considerations that determine whether investment in a biomarker-based study design should be made. Three interrelated statistical considerations are the anticipated effect size, prevalence of Dx +ve group, and accuracy measures of the diagnostic test. Hypothetical examples and simulations will be utilized to examine the effects of the key statistical considerations and demonstrate how results of such analyses can assist in deciding if biomarker-based designs should be preferred.

Can Traditional Pharmacokinetic Nonlinear Models Be Replaced With Random Effects Linear Models?

✦Francisco J Diaz, The University of Kansas Medical Center, Department of Biostatistics, Mail Stop 1026, 3901 Rainbow Blvd., Kansas City, KS 66160, *fdiaz@kumc.edu*

Key Words: Random effects, Linear models, Nonlinear models, Drug-drug interactions, Drug dosage individualization, Pharmacokinetics

Traditional pharmacokinetic analyses use nonlinear models of drug plasma (or blood) concentrations. However, published empirical findings show that random effects linear models may provide accurate representations of phase III and IV steady-state pharmacokinetic data, and may be useful for dosage computations. In addition, experienced statisticians know that linear models are much easier to build and fit than nonlinear models. Another good reason for using linear models in the analysis of steady-state pharmacokinetic data is the sparse sampling designs that usually characterize phase III and IV studies, which impede a clear determination of absorption parameters. In this paper, we describe successful applications of random effects linear models to pharmacokinetic research, particularly to drug-drug interaction studies. We also describe new, published developments that show that random effects linear models may provide a solid theoretical framework for drug dosage individualization in chronic diseases. In particular, in-

dividualized dosages computed with these models may produce better results than dosages computed with some methods routinely used in therapeutic drug monitoring.

A Range Of Practical Issues In Pediatric Early Phase Trials: Dosing Approaches And Empirical Vs. Model-Based Phase I Designs

✦Arzu Onar-Thomas, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, *arzu.onar@stjude.org*

Key Words: Continual Reassessment Method, Phase I Trial, MTD

Pediatric Phase I trials can differ from their adult counterparts and pose unique challenges. One specific issue is with respect to dosing oral agents, which is often based on body surface area in pediatrics. With fixed pill sizes and without availability of pediatric formulations, using such agents in a Phase I setting requires special care for ensuring safety and for estimating the maximum tolerated dose. Various dose-finding designs are available for Phase I trials, some empirical and others model-based. Here we will present extensive simulation results that compare the performance of empirical designs, specifically the Rolling-6 and the Traditional Method, to each other and to the continuous reassessment method with respect to accuracy, sample size and toxicity. The advantages/disadvantages of using each design will be highlighted in the context of pediatric Phase I oncology trials.

Statistical Methods And General Principles Of Exposure-Response Analysis

✦Rui Tang, Amgen inc, one center amgen drive, thousand oaks, CA 91320, *rtang@amgen.com*; Erik Rasmussen, Amgen Inc; Hongjie Deng, Amgen inc; Lisa Hendricks, Amgen Inc; Mike Hale, Amgen Inc; Li Chen, Amgen Inc

Key Words: PK, PD, exposure response analysis, protocol designs

Pharmacokinetics (PK) characterizes the absorption, distribution, metabolism, and elimination properties of a drug, while pharmacodynamics (PD) defines the physiological and biological response to the administered drug. Exposure-response analysis seeks to explore and characterize the relationship between these processes. To understand the relationship between drug exposure and efficacy (or safety) response, extensive exposure-response analyses using data from a phase 2 study were conducted. The analyses impacted dose selection for the phase 3 protocol designs. In this presentation, we will summarize the statistical methods and principles for exposure-response analyses. Issues and limitation of these analyses will be indentified and discussed. Experiences with cross-functional collaboration will be shared.

A Novel Statistical Methodology For Finding A Clinically Meaningful Threshold

✦Xin Fang, US FDA, 10903 New Hampshire Ave., Silver Spring, MD 20903, *xin.fang@fda.hhs.gov*; Mahboob Sobhan, US FDA

Key Words: clinically meaningful threshold, negative predicted value, positive predicted value

A clinically meaningful threshold (CMT) is defined as an efficacy cutoff value. If, for example, a patient's efficacy response is below the CMT, the patient has not experienced a clinically meaningful improvement. Estimates of CMT are based on the subjective perspectives of investigators. Currently, there is no statistical method to estimate the CMT. The authors have developed and investigated a novel two-step statistical methodology for locating the CMT. This method provides a non-traditional approach for establishing the theoretical (or probability) background for each of the four components of a receiver operating characteristic curve. Consequently, both positive predicted value and negative predicted value are used in investigating the threshold. Simulation results will be used to illustrate some of the properties of the derived CMT estimators.

Statistical Inference For Dynamic Systems Governed By Differential Equations With Applications To Toxicology

✦Siddhartha Mandal, UNC Chapel Hill, 416 W Cameron Ave, Apt 6, Chapel Hill, NC 27516, *sid.stat.iitk@gmail.com*; Pranab K Sen, University of North Carolina at Chapel Hill; Shyamal D Peddada, National Institute of Environmental Health Sciences

Key Words: Differential equations, physiologically based pharmacokinetic, basis expansions

Stochastic and deterministic differential equations are used to describe a wide variety of biological and physiological phenomena. For example, in Physiologically based pharmacokinetic(PBPK) models, differential equations explain the absorption, distribution, metabolism and excretion (ADME) of a compound in the human or animal body. Usual approaches for parameter estimation in such situations include non-linear least squares and Bayesian hierarchical modeling. However, a common challenge with these problems is the lack of explicit equations/models that relate response variable to the explanatory variables. Recent functional data analysis methods indicate the use of basis functions to bypass this problem. This talk focuses on estimation and inference of the model parameters, taking into account the variability within and between multiple subjects, while exploiting the structure implied by the system of differential equations. Large sample behavior of the parameter estimates are also explored. Application of the methods are shown using simulated and real life data on compartmental and state space models.

The Real Implication Of Incorporating Pk Into Dose-Escalation Trials

✦Cheng Zheng, Novartis Pharmaceuticals Corporation, 180 Park Ave., Florham Park, 07932, *zhengcheng@gmail.com*; Lu-May Chiang, Novartis Pharmaceuticals Corporation

Key Words: Bayesian hierarchical model, dose-escalation, pharmacokinetic, Phase I

Extensive research has been conducted using Bayesian techniques for modeling dose-toxicity relationship in oncology dose-escalation trials. Many have attempted to integrate pharmacokinetic (PK) information, such as drug exposure measured by area under the curve (AUC), into the model. However, given these proposals were rarely formally evaluated, the implication of such practice in real trials remains unknown. We are proposing a Bayesian hierarchical model considering both the PK exposure-toxicity curve and the relationship between dosage and PK. This model can be fitted via Gibbs sampling and is very flexible in addressing safety concerns. We have studied this model in details based on data from 5 past or ongoing Novartis Phase I trials with large number of enrolled patients and well-collected PK profiles. After comparing

the new model with the standard Bayesian model for in-house dose-escalation trials, we propose conditions under which incorporating PK information is beneficial during dose-escalation in terms of both maximum tolerated dose (MTD) estimate accuracy and safety profile.

352 Asymptotics in Time Series

IMS

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Asymototic Theory For General Multivariate Garch Models

✦Weibin Jiang, The University of Western Ontario, Western Science Centre - Room 228, The University of Western Ontario, London, ON N6A 5B7 Canada, *wjiang@stats.uwo.ca*; Hao Yu, The University of Western Ontario; Reg Kulperger, The University of Western Ontario

Key Words: Multivariate GARCH, QMLE, Identifiability, Ergodicity, Strong Consistency, Asymptotic Normality

Generalized autoregressive conditional heteroscedasticity (GARCH) models are widely used in financial markets. Parameters of GARCH models are usually estimated by the quasi-maximum likelihood estimator (QMLE). In recent years, economic theory often implies equilibrium between the levels of time series, which makes the application of multivariate models a necessity. Unfortunately the asymptotic theory of the multivariate GARCH models is far from coherent since many algorithms on the univariate case do not extend to multivariate models naturally. This paper studies the asymptotic theory of the QMLE under mild conditions. We give some counterexamples for the parameter identifiability result in Jeantheau (1998) and provides a better necessary and sufficient condition. We prove the ergodicity of the conditional variance process on an application of theorems by Meyn and Tweedie (2009). Under those conditions, the consistency and asymptotic normality of the QMLE can be proved by the standard compactness argument and Taylor expansion of the logarithm of the score function.

Fixed-Smoothing Asymptotics For Time Series

✦Xianyang Zhang, University of Illinois at Urbana-Champaign, Department of Statistics, 725 South Wright Street, Champaign, IL 61820, *zhang104@illinois.edu*; Xiaofeng Shao, University of Illinois at Urbana-Champaign

Key Words: Fixed-smoothing, generalized method of moments, High order expansion, Long run variance

This paper proposes a wide class of estimators for estimating the asymptotic covariance matrix of the GMM (generalized method of moments) estimator in the stationary time series models. The proposed estimator is general enough to include the traditional heteroskedasticity and autocorrelation consistent covariance estimatior and some recent developed estimators, such as cluster covariance estimator and projection-based covariance estimator, as special cases. Under the framework of Gaussian location model, we derive a high order expansion for the corresponding Wald statistic when the underlying smoothing parameter is held fixed. Specifically, we show that the error rejection probability is of order $O(1/T)$, where T is the sample size, and derive the leading term in the expansion under the fixed-smoothing asymptotics. Furthermore, we propose a novel bootstrap method, called Gaussian dependent bootstrap, and show that the bootstrap based inference is more accurate than the first order approximation.

Asymptotic Theory Of Fractal Index And Scale Parameter Of Irregularly Observed Gaussian Field

✦Myoungji Lee, University of Chicago, 5734 S. University Avenue, department of statistics, chicago, IL 60637 usa, *myoungji@uchicago.edu*; Michael Stein, University of Chicago

Key Words: fractal dimension, increments, variogram, Gaussian random process, least squares estimation, fixed domain asymptotics

The fractal dimension is a scale invariant measure of quantifying how rough or smooth a curve or surface is. Assuming that the process is stationary isotropic Gaussian and observations are even, the variogram based estimation of the fractal dimension is consistent and follows normal or Rosenblatt distribution with slower speed than root n depending on the smoothness of the field. Uniform root n convergence to normal distribution is achieved by using the high order increments in a lattice. This paper extends the work by allowing uneven observations, which is far more natural than evenly spaced data. We first expand the concept of the increment to accommodate irregularity of observations. Then using the squared increments, the least squares estimators of the fractal index and the scale parameter of the covariance function will be proposed and their consistency and asymptotic normality will be shown. Under fixed domain asymptotics, the same rates of convergence are achieved for unequally spaced data as for equally spaced data under some regularity assumptions on the amount of irregularity of locations of observations.

Asymptotics Of Markov Order Estimators For Infinite Memory Processes

✦Zsolt Talata, Department of Mathematics, University of Kansas, 1460 Jayhawk Boulevard, Snow Hall, Room 405, Lawrence, KS 66045-7523, *talata@math.ku.edu*

Key Words: Markov order estimator, information criterion, divergence rate, asymptotics, ergodic process, infinite memory

For finite-alphabet stationary ergodic processes with infinite memory, Markov order estimators that optimize an information criterion over the candidate orders based on a sample of size n are investigated. Three familiar information criteria are considered: the Bayesian information criterion (BIC) with generalized penalty term yielding the penalized maximum likelihood (PML), and the normalized maximum likelihood (NML) and the Krichevsky-Trofimov (KT) code lengths. A bound on the probability that the estimated order is greater than some order is obtained under the assumption that the process is weakly non-null and alpha-summable. This gives an O(log n) upper bound on the estimated order eventually almost surely as n tends to infinity. Moreover, a bound on the probability that the estimated order is less than some order is obtained if the decay of the continuity rate of the weakly non-null process is in some exponential range. This implies that then the estimated order attains the O(log n) divergence rate eventually almost surely as n tends to infinity.

Asymptotics For Time-Varying Autoregressive Processes

✦SREENIVAS KONDA, Temple University, 345 Speakman Hall, Fox School of Business, Philadelphia, 19122, *konda@temple.edu*

Key Words: Time series, Nonstationary, Locally stationary, Local likelihood, Strong mixing, EEG

A theoretical framework is proposed for the asymptotics of time-varying autoregressive parameters' estimates in time domain. First we simplify the problem by fitting the time dependent parameters of these nonstationary processes by local polynomial models and then estimate the parameters of these approximate models by weighted least squares method. The asymptotics of the proposed estimators are derived and studied under strong mixing conditions. The developed asymptotics are further investigated and compared using Gaussian errors and Fisher Information Matrix. Local linear models seem to exhibit some optimal properties if bias is corrected or minimized. One useful outcome of this research is to apply minimum biased kernel smoothers for larger bandwidth windows. Such method expected to obtain small standard error values for the estimators, hence the smaller MSE. This research will be presented and explained using EEG data.

Results On The Maximum Of Time-Dependent Regularly Varying Variables

✦Chris O'Neal, University of Georgia Department of Statistics, 1050 South Lumpkin Street, Apt 206, Athens, GA 30605, *chrisoneal2718@gmail.com*; William P. McCormick, University of Georgia Department of Statistics; Lynne Seymour, The University of Georgia

Key Words: Regular Variation, Extreme Value Theory, Annual Maximum

We derive an asymptotic distribution for the maximum of a time-dependent series of regularly varying random variables. The choice of normalizing constants is discussed, and simulation results illustrate the effectiveness of this approximation. We also test the formula on the annual maximum gage heights from the Peachtree Creek in Atlanta.

Max Autoregressive Processes: Approximations And Time Series Models With Log-Positive Alpha Stable Noises And Hidden Max Gumbel Shocks

✦Bin Zhu, UW-Madison, , *binzhu@stat.wisc.edu*; Zhengjun Zhang, University of Wisconsin; Philippe Naveau, Laboratoire des Sciences du Climat et de l'Environnement

Key Words: Extreme Value Theory, Dependence, Gumbel Distribution, Autoregressive Model

Max-autoregressive (MAR) processes and moving maxima (MM) processes are naturally adapted from linear autoregressive (AR) processes and moving average (MA) processes in modeling clustered maxima in time series. Yet, applications of MAR processes and MM processes are still sparse due to some di±culties of statistical parameter estimation and some abnormality of the processes, basically that ratios of observations can take constant values. The objective of this present work is to introduce a new model that is closely related to MAR processes and is free of the aforementioned abnormality. A logarithm transformation of the new model leads to time series models with log-positive alpha stable noises and hidden max Gumbel shocks. Theoretical properties of new models are derived.

353 Methods of Assessment ■●

Section on Statistical Education, Section on Teaching of Statistics in the Health Sciences

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

A Study on Perceptions of Board Members and Superintendents on Issues Affecting Student Learning Outcomes

✦Dai-Trang Le, Iowa State University, 169 University village, Unit G, Ames, IA 50010, *daitrang@iastate.edu*

Key Words: proportional odds, logistic regression, board members and superintendents, student learning outcomes

Abstract In question #21 of the questionnaire for the National School Board and Superintendent Survey conducted in 2008, respondents were asked to rate on the scale of 0 to 4 the important level of each of the 11 selected criteria that were considered issues affecting student learning outcomes. This project investigates factors that affect these board members and superintendents' rating scores on those 11 issues. We use proportional odds logistic regression method to analyze the data. The results will be revealed and discussed.

The Effect Of Structured Interviews On Predicting The Success Of Nursing Students

✦Renjin Tu, Columbus State University, Dept. of Math, Columbus State University, Columbus, GA 31907, *tu_renjin@ColumbusState.edu*

Key Words: correlation analysis, multiple regression, logistic regression

Due to the increasing need for nurses and the limited faculty resources available, it is necessary to determine the essential criteria to be used in the selective admission process. There is abundant research showing the predictable value of standardized tests and grade point averages(GPA), but there have been very few studies conducted that are concerned with the ability of individual interviews to predict outcomes. This study will address the gap in the literature related to the effects of adding structured interviews to cognitive-based admissions processes. The performance in college admission standardized test scores, specialized pre-nursing standardized test scores, prerequisite course grades, gender, individual interviews scores, and GPA of 209 students are analyzed by correlation, regression and logistic analysis. The main results show that individual interview score is not significantly associated with GPA. However, individual interview score, prerequisite course grade, and gender are significantly associated with the probability of the completeness of the program.

Use of Factor Analysis in Assessments of Clinical Teaching Evaluations

✦Jay Mandrekar, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, *mandrekar.jay@mayo.edu*

Key Words: Factor Analysis, Latent Factors

Factor analysis is a generic term for a family of statistical techniques concerned with the reduction of a set of observable variables in terms of a small number of latent factors. It has been developed primarily for analyzing relationships among a number of measurable entities (such as survey items or test scores). The underlying assumption of factor analysis is that there exists a number of unobserved latent variables (or "factors") that account for the correlations among observed variables, such that if the latent variables are partialled out or held constant, the partial correlations among observed variables all become zero. In other words, the latent factors determine the values of the observed variables. Factor analysis has been widely used, especially in behavioral sciences, to assess the construct validity of a test or a scale. The focus of this talk is to provide an introduction to factor analysis in the context of research projects from Medical Education that involve clinical teaching evaluations, for example, resident-teacher evaluations, resident's reflection on quality improvement etc.

Bayesian Hierarchal Models Of Final Exam Grades In Statistics Classes, A Two-University Study

✦Lawrence V Fulton, Texas State University, 601 University Drive, San Marcos, TX 78108, *lf25@txstate.edu*; Rasim Musal, Texas State University; Lana Ivanitskaya, Central Michigan University; Salma Haidar, Central Michigan University; Carl Lee, Central Michigan University

Key Words: statistics, Bayesian, zero-inflated, education, hierarchical

In this two university, IRB-approved study of business statistics courses, we use Bayesian hierarchical modeling to search for the effects of pre-existing ability, demographics, manipulated homework certification standards, "cramming," and other covariates on the probability that a student achieves a particular grade. The significance of our research is that we are able to determine probabilistically the grade that students should expect given fixed covariate values. This information is valuable to both professors and students in realistically appraising likely performance. Our model is nonlinear in parameters and reflects a zero-inflated distribution, accounting for the population of students who did not have an economically rational reason to take the final examination. We identify the distinct number of student grade groups and employ mixed beta distributions in modeling the examination scores. By using Bayesian hierarchical modeling, we are able to obtain the examination grade probability distribution that any student "i" will obtain given covariate values. As part of our research, we conduct cross-validation and evaluate model fit.

Items Reduction in Course Evaluation Survey Questionnaire: An Analytical Appraisal

Abdullah Al Rubaish, University of Dammam; Lade Wosornu, University of Dammam; ✦Sada Nand Dwivedi, University of Dammam, Deanship of Quality & Academic Accreditation, Building 10, Main University Campus, Dammam, 31451 Saudi Arabia, *dwivedi7@hotmail.com*

Key Words: Item reduction, Course evaluation survey, Factor analysis, non-co-colinearity, inter-correlation, overall reliability

The University of Dammam (UoD) is developing robust evidence required for academic accreditation. One of the important tools is course evaluation survey (CES) questionnaire. This study aims to employ theoretical, pragmatic and analytical considerations to optimize the number of items in this questionnaire while preserving its theoretical structure and analytical power. For exploratory analysis, the currently available CES data in four colleges (nursing, medicine, dentistry and engineering) were used. The college-specific number of courses ranged between 1 to 15 and the number of surveyed students from 32-113. Evidence of co-linearity among the items paved the way to systematically reduce the number of items in the original questionnaire. The analytical methods included consideration of data collected on ordinal scale as internal scale, correlation matrices and factor analysis. In spite of the modest sample size, three desirable characteristics of a questionnaire (non-co-linearity among items, inter-correlation among items, and overall reliability) showed either improvement or more consistency in the shortened questionnaire than that in original questionnaire. In summary, the origin

Accumulation of Equating Error and Random Walk

✦Hongwen Guo, Educational Testing Service, Rosedale Rd-MS 02-P, Princeton, NJ 08541 US, *hguo@ets.org*; Jinghua Liu, ETS; Neil Dorans, ETS

Key Words: random walk, test equating, sampling error

In testing industry, equating is a procedure to adjust for the small differences in test form difficulty so that scores obtained from different forms are interchangeable. However, all equating is subject to equating error, either systematic or random. An equating process can maintain the score scale for some time; the cumulative effects of changes might result in scores at one time being not comparable with scores at a later time. It is crucial to solve this problem in testing industry. When a series of individual equatings are concatenated over time, the equating error is accumulated as a random walk which will lead to shifts, even explosion in score scales. A comparison between single- and multiple-link equating is investigated. It was shown that multiple-link equating procedure can slow down the explosion of the error in practice.

Statistical Analysis Of An Educational Research Study

✦Carolyn Bradshaw Morgan, Hampton University, Queen and Tyler Streets, Hampton, VA 23668, *carolyn.morgan@hamptonu.edu*; Anne Pierce, Hampton University; Spencer Baker, Hampton University; Arun Verma, Hampton University; Morris Morgan, Hampton University; Vitali Khaikine, Hampton University

Key Words: data analysis, assessment, education curriculum, statistical education

Statistics is indeed an all-encompassing discipline. Good statistical data analysis is critical to evaluating the success of many educational research studies. A research study is being conducted to investigate the impact of a new instructional format on student learning outcomes in calculus. One of the goals of the study is to assess the importance of faculty mentors and role models on the performance of science, technology, engineering and mathematics (STEM) majors, especially members of underrepresented groups. The study addresses the STEM student population enrolled in the gate-keeper Calculus I course. Some

of the calculus classes have received the standard classroom instruction while other sections have received the FORCE (financially oriented research calculus experience) classroom instruction. All sections of the course receive a pre- and post-calculus readiness test. All students take a standard final exam and a grade of "C" or higher is considered passing for all students. A discussion of the statistical analysis techniques and results to date will be presented.

354 Generalized Linear and Partial Linear Models ●

International Chinese Statistical Association, International Indian Statistical Association

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Selection Of Variables And Smoothing Parameters In Semiparametric Additive Models

✦Ji Luo, Zhejiang University of Finance and Economics, Genshanfuju 9-2-1001, Huancheng North Road, Hangzhou, International 310004 China, *luoji0409@gmail.com*; Shurong Zheng, Northeast Normal University

Key Words: additive partial linear model, selection of variables, degree of smoothing, CV, BIC

The paper propose a new semiparametric model, the accelerated failure time additive partial linear model, which allows the response variable to depend on more than one nonparametric covariates and linearly on the remaining variables. The emphasis of the paper is placed on investigating a new methodology for simultaneously selecting variables and degrees of smoothing in the proposed semiparametric additive models. The procedure obtains the sequence of covariates according to their importance in order to decide a particular covariate enters the model in linear part, or in nonparametric part or is removed from the model, where CV or BIC selector can be used to select the tuning parameters. Some simulation studies have been conducted to show the good performance of the procedure.

Robust Empirical Likelihood Inference For Generalized Partial Linear Models With Longitudinal Data

✦Yang Bai, Shanghai University of Statistics and Management, 777 Guoding Road, Shanghai, International 200433 China, *statbyang@mail.shufe.edu.cn*; Guoyou Qin, Fudan University; Zhongyi Zhu, Fudan University

Key Words: B-spline, efficiency, empirical likelihood, generalized estimating equations, longitudinal data, robustness

In this paper, we propose the robust empirical likelihood (REL) inference for the parametric component in a generalized partial linear model (GPLM) with longitudinal data. We make use of bounded scores and leverage-based weights in the auxiliary random vectors to achieve robustness against outliers in both the response and covariates. Simulation studies demonstrate the good performance of our proposed REL method, which is more accurate and efficient than the robust generalized estimating equation (GEE) method (He et al. 2005, Journal of the American Statistical Association 100, 1176-1184). The proposed robust method is also illustrated by analyzing a real data set.

Prediction Accuracy Of Linear Models For Paired Comparisons In Sports

✦Victor Chan, Western Washington University, 202 Bond Hall, Department of Mathematics, Bellingham, WA 98225, *victor.chan@wwu.edu*

Key Words: Bradley-Terry model, Thurstone-Mosteller model, binomial distribution, negative binomial distribution, poisson distribution

Linear models for paired comparisons, the Bradley-Terry model and the Thurstone-Mosteller model in particular, are widely used in sports for ranking and rating purposes. By their formulation, these models predict the probability that a player or team defeats another if the playing strengths of the players or teams are known. In this paper, we investigate the prediction accuracy of the two linear models by using them to describe three simple theoretical games which mimic actual sports and whose winning probability, given the playing strength of each player, can be expressed explicitly. A theoretical result is presented, which provides the basis of a linearization method that enables these games to be represented by linear models. The predicted winning probabilities from the linear models are then compared to the actual ones. Comparisons are also made in prediction accuracy between the Bradley-Terry model and the Thurstone-Mosteller model.

A Simple And Efficient Reparameterization For Mixed-Effects Models

✦Guangxiang Zhang, State University of New York at Stony Brook, NY 11764 USA, *georgexzh@gmail.com*; John J. Chen, Stony Brook University Medical Center

Key Words: Mixed-effects or multilevel models, Convergence rate, Collinearity between random-effects, Centering, Optimal linear transformation, Random slope

Linear mixed-effects model has been widely used in hierarchical and longitudinal data analyses. In practice, the fitting algorithm can fail to converge due to boundary issues of the estimated random-effects covariance matrix G. Current available algorithms are not computationally optimal because the condition number of G is unnecessarily increased when the random-effects correlation estimate is not zero. The traditional mean centering technique may even increase the random-effects correlation. To improve the convergence of data with such boundary issue, we propose an adaptive fitting (AF) algorithm using an optimal linear transformation of the random-effects design matrix. The AF algorithm can be easily implemented with standard software and be applied to other mixed-effects models. Simulations show that AF significantly improves the convergence rate, and reduces the condition number and non-positive definite rate of the estimated G, especially under small sample size, relative large noise and high correlation settings. One real life data for Insulin-like Growth Factor (IGF) protein is used to illustrate the application of this algorithm implemented with software package R (nlme).

Statistical Strategy For Eqtl Mapping Using Rna-Seq Data

✦Wei Sun, University of North Carolina, Chapel Hill, 432 Edisto Ct, Chapel Hill, NC 27514 United States, *weisun@email.unc.edu*

Key Words: RNA-seq, allele-specific expression, isoform, generalized linear model, penalized regression, alternative splicing

RNA-seq is going to replace gene expression microarray in the near future for genome-wise assessment of the transcriptome. In addition to be more accurate and more sensitive, RNA-seq also provides new information that is not available from microarray, and novel statistical methods are needed to fully explore these new information. We will introduce two statistical approaches that aim to explore two types of characteristics of RNA-seq data for eQTL (gene expression quantitative trait loci) mapping. First, we develop a likelihood-based approach for eQTL mapping using both total expression and allele-specific expression measurements. Modified generalized linear models are separately developed for total expression and allele-specific expression, and then the two models are joined together by a sharing parameter. Secondly, we will discuss a penalized regression approach for mapping the variation of the isoforms proportions of a gene, or in other words, the genetic basis of alternative splicing.

Seasonality Analysis Of Time Series In Partial Linear Models

✦Qin Shao, University of Toledo, Math Department Mailstop 942, Toledo, OH 43606, *qin.shao@utoledo.edu*

Key Words: Partial Linear Model, Mean Integrated Squared Error, Local Linear Estimation, Seasonal Component

Seasonality analysis is one of the classic topics in time series. This paper studies techniques for seasonality analysis when the trend function is unspecified. The asymptotic properties of the semiparametric estimators are derived. An estimation algorithm is provided. The techniques are applied to making inference for the monthly global land-ocean temperature anomaly indexes.

Estimating Variance Components And Variance Partition Coefficients On The Inverse Link Scale For Cross-Classified Random Effects Models

✦Brian R Gray, US Geological Survey, 2630 Fanta Reed Rd, La Crosse, WI 54603, *brgray@usgs.gov*

Key Words: hierarchical models, Laplace estimation, logit-normal models, multilevel models, variance partition coefficient, variance components

Studies of the estimation of variance components (VCs) and relative VCs (variance partition coefficients, VPCs) on inverse link scales from generalized linear mixed model (GLMM) estimates have been limited to outcomes from nested designs. Also, the influence of GLMM analytical method on VC and VPC estimation has received limited attention. I propose an approach for calculating VCs and VPCs on inverse link scales from outcomes from two-way cross-classified random effects designs, and evaluate this method using Monte Carlo simulations of grouped binomial data. GLMMs were fitted using first-order marginal and penalized quasi-likelihood (PQL1), the REML analogue of PQL1 (RPQL1), Laplace estimation and Markov chain Monte Carlo. Bias and precision in VC and VPC estimates improved as group (cluster) size for both random main effects increased and, for a given main effect, when the number of groups associated with the alternate main effect increased. The influence of estimation method on bias and precision of VC and VPC estimates was typically slight when numbers of groups for both main effects were 10 and 20 but when numbers of groups equaled 5 were best under RPQL1.

355 Advanced Applications of Statistics in Marketing ■

Section on Statistics and Marketing, Section for Statistical Programmers and Analysts

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Optimization Of Taxi Site Configuration Based On Historical Statistic Demand And Forecasting Trip Amount

✦Wei Zhao, IBM Research-China, Diamond Building, Zhongguancun Software Park #19,, Beijing, International 100193 China, *wzhaow@cn.ibm.com*; Xinxin Bai, IBM Research-China; Jinyan Shao, IBM Research-China; Ming Xie, IBM Research-China; Wenjun Yin, IBM Research-China; Jin Dong, IBM Research-China

Key Words: Optimization of taxi site configuration, historical statistic demand, forecasting trip amount, constrained optimization, penalty function

A novel approach to optimize the taxi site configuration based on historical statistic demand and forecasting trip amount is presented. Considering the limitation of resources, we combine the optimization for configuration of single taxi site with global planning of site configuration in a whole city. The problem for configuration of single taxi site is modeled as constrained optimization problem, which can be solved by combination of penalty function and one-dimensional search methods based on the historical statistic data of its net taxi demand. The global planning of site configuration in a whole city is conducted by imputation method based on two types of data for each site, including surrounding traffic information and forecasting data of surrounding citizens' trip amount by taxi. Finally, an example with results is given to show the effectiveness of our approach.

Bank Branch Channel Preference Analysis

✦Jin Sun, IBM Research China, Building 19# Zhongguancun Software Park,, 8 West Dongbeiwang Road, Haidian District, Beijing, International 100094 China, *sunjinsj@cn.ibm.com*; Bin Zhang, IBM Research China; Ming Xie, IBM Research-China

Key Words: branch channel preference, information gain, decision tree, a unified feature view

ATM is an important and frequently used service channel within bank branches. How to determine the number of ATMs within a bank branch, however, is a difficult and challenging problem. To address this problem, the first step is to determine customers' preference on ATMs. In this paper, the transaction records of over ten thousand clients of a typical bank branch are investigated. Such data is used to construct a

unified feature view of customers including basic information, transaction behavior, and product preference. The information gain of each feature with respect to the preference on ATMs is calculated to screen out a small subset of features with larger information gain. Based on the samples and the selected features, a decision tree is derived with a prediction accuracy of 85%.

Statistical Analysis And Forecasting Of The Business Load Of A Typical Banking Service Node Based On Season Index Forecasting Method

✦Ming Xie, IBM Research-China, 100193 China, *xieming@cn.ibm.com*; Jin Sun, IBM Research China; Bin Zhang, IBM Research China

Key Words: business Load forecasting, bank branch, season index forecasting

The future business load is the main uncertainty challenging the labor scheduling of a service-oriented company, for example, a branch bank maintaining a service network which contains several subsidiary service nodes at different places. Underestimation or overestimation of the future business load will lead to loss of business or labor redundancy, and thus loss of customer satisfaction or unnecessary labor cost, respectively. So business load forecasting is important as the prerequisite for developing a suitable labor reserving and scheduling plan. In this study, we focused on the business load forecasting for a bank service node. Through statistically analysis on the data from some bank branch, we found certain characteristics from its historical business load data. Taking advantage of such characteristics, we developed a forecasting method based on the widely-used season index forecasting method. Tests show that our method works well for the bank service node investigated. In fact, the idea of the method, which is what this paper mainly wants to deliver, can be used to develop forecasting methods for other similar scenarios.

Statistical Analysis And Optimization Of Business Load Balancing Of A Bank Branch Service Network

✦Bin Zhang, IBM Research China, Building 19# Zhongguancun Software Park,, 8 West Dongbeiwang Road, Haidian District, Beijing, International 100094 China, *zbin@cn.ibm.com*; Ming Xie, IBM Research-China; Jin Sun, IBM Research China

Key Words: business load balancing, integer programming, ILOG CPLEX, optimization

A bank branch usually maintains a service network which contains several subsidiary service nodes at different places, and has a number of clerks working for that network. Developing a suitable scheduling table, i.e., determining for each clerk which days he should work and which node he should work at, can be a tough work if only by rule of thumb, especially when the number of nodes and clerks are large. Through statistical analysis on the data from some bank branch, we found that the current scheduling table of the bank branch is not reasonable in the sense that it may result in imbalanced business load distribution, e.g., the average business load per clerk differs greatly in different nodes in the same day, or in different days in the same node. To reduce the imbalance of workload, we developed an integer programming model of scheduling, taking into account realistic constraints and objectives raised by the bank branch. We used IBM ILOG CPLEX to solve this problem and obtained a much balanced scheduling table.

Smart Robots At Home: The Factors Affecting Consumer'S Acceptance Of Self-Service Technology Products

✦Nai-Hua Chen, Chienkuo Technology University, 500 Taiwan, *nhc@cc.ctu.edu.tw*; Stephen C.T. Huang, National Kaohsiung First University of Science and Technology

Key Words: echnological readiness, technology acceptance model, perceived risks; structural equation modeling

As the technology advances rapidly, a variety of self-service technologies (SSTs) have been introduced to the market. This study proposes a model to identify the factors that influence consumer's acceptance of the self-service products. Drawing on the theories of technological readiness (TR), technology acceptance (TAM), theory of planned behavior (TPB), perceived risks and cognitive age, the proposed model posits relationships among consumer's adoption intention for family-use self-service products and its antecedents. The focal product selected is the family oriented smart robots which are self-service oriented in the daily life. Because of the mature technology in robot development, some companies, for instance Google, Intel, Microsoft, SONY and HONDA, are now devoted in launching the robots for family uses, such as house keeping, health cares, the security management and family interactions. Hypotheses are tested using a structural equation modeling method (SEM-PLS).

A State-Space Approach To Capturing Market Dynamics For Frequently Purchased Products

✦Eiji Motohashi, The Graduate University for Advanced Studies, 10-3 Midori-cho, Tachikawa, International 1908562 Japan, *eiji.motohashi@gmail.com*; Tomoyuki Higuchi, The Institute of Statistical Mathematics

Key Words: State-space approach, Multinomial logit models, Brand choice, Marketing dynamics

Marketers are often interested in how consumer preferences and/or promotion effects vary over time. Previous scholars have proposed many approaches to model such market dynamics in consumer brand choices. Parameters in a choice model may vary over time because of a variety of reasons such as change in market conditions. In this study, we develop a choice model that includes time-varying parameters (state variables) based on the state-space approach and show how we incorporate both invariant individual heterogeneity and variant marketing factors into the model. In empirical analysis, we use the particle filter method to estimate state variables and unknown parameters, and apply our methodology to scanner panel data in instant coffee category, which have been gathered at a super market in Japan. We find that the proposed model can obtain more appropriate inferences about brand choices than models that ignore time variation in parameters.

Estimating Planned Sales Call Frequencies With Incomplete Information Using The Em Algorithm

✦Lan Ma Nygren, Rider University, 2083 Lawrenceville Road, Lawrenceville, NJ 08648, *lnygren@rider.edu*; Lewis Coopersmith, Rider University

Key Words: EM algorithem, Incomplete information, Multinomial cell probabilities, Sales call frequencies, Diary Survey

We consider estimating planned sales call frequencies of a selling company with incomplete information caused by short recording durations in diary surveys. For practical reasons, it is necessary to keep the recording period short. Missing data occur when the recording period is not long enough to include observations with low call frequencies. We derive the maximum likelihood estimators of the multinomial cell probabilities for the planned sales call frequencies using the expectation maximization (EM) algorithm. We show that the EM algorithm estimators have good asymptotic properties in terms of both bias and mean squared error (MSE) and are more accurate and reliable than the estimators obtained by the naive approach of treating the absence of a sales call as a non-called on respondent (i.e., zero frequency). The effect on the estimators when the number of frequency classes increases is also investigated.

356 Method for Ecological Data ■●

Section on Statistics and the Environment, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

A Multivariate Regression Tree Approach To Defining Population Spatial Units Using Frequencies Of Individual Characteristics And Time Series Of Abundance

✦Cleridy E. Lennert-Cody, Inter-American Tropical Tuna Commission, 8604 La Jolla Shores Drive, La Jolla, CA 92037-1508, *clennert@iattc.org*; Mark N. Maunder, Inter-American Tropical Tuna Commission; Alexandre Aires-da-Silva, Inter-American Tropical Tuna Commission; Mihoko Minami, Keio University

Key Words: multivariate regression tree, Kullback-Leibler divergence, population management unit, stock structure

A critical aspect of the assessment of population status is identification of areas containing relatively independent populations or individuals with similar characteristics. We present an approach for defining spatial units that combines results of multivariate regression tree analyses of spatial structure in frequency distributions of individual characteristics and in temporal trends in abundance. We illustrate this approach with length-frequency and catch-per-unit-effort data for bigeye tuna (Thunnus obesus) from the Japanese longline fisheries in the eastern Pacific Ocean. The approach proves useful as a means of quantitatively defining dominant spatial structure, and provides options for incorporating non-quantitative subject-matter knowledge into the choice of spatial assessment units. Although our application is focused on population management, this approach could be applied to any system for which data describing the system dynamics were available in the form of time series and frequencies of characteristics.

Modeling Animal Abundance And Detection With A Hierarchical Catch-Effort Model

✦Katherine St. Clair, Carleton College, 1 North College St., Northfield, MN 55057 USA, *kstclair@carleton.edu*; John Giudice, Minnesota Department of Natural Resources; Eric Dunton, Shiawassee National Wildlife Refuge

Key Words: abundance estimation, removal sampling, harvest model, catch-effort, Bayesian analysis, hierarchical models

A hierarchical modeling framework can produce estimates of animal abundance and detection from replicated removal counts taken at different locations in a region of interest. A common method of specifying variation in detection probabilities across locations or replicates is with a logistic model that incorporates relevant detection covariates. As an alternative to this logistic model, we propose using a catch-effort model to account for heterogeneity in detection when a measure of removal effort is available for each removal count. We model the probability of detection as a nonlinear function of the removal effort and a catchability parameter that can vary spatially and temporally. Simulation results demonstrate that Bayes estimates from our model are effective estimates of abundance and catchability. We also found that our catch-effort model fits better than logistic models when estimating wild turkey abundance using harvest and hunter counts collected by the Minnesota Department of Natural Resources.

Parameter Estimation In A Stochastic Nonlinear System Of Difference Equations

✦Sam Woolford, Bentley University, 175 Forest St, Waltham, MA 02452, *swoolford@bentley.edu*; Mihaela Predescu, Bentley University; Norm Josephy, Bentley University

Key Words: 2-stage model, MLE estimation evaluation, nonlinear stochastic model, population model

This paper considers a stochastic version of a deterministic, two-stage nonlinear discrete population model that has been used to explain the oscillatory behavior observed in some fish populations. Under certain distributional assumptions, the various model parameters can be estimated using maximum likelihood estimation (MLE). However, in practice, it is very difficult to determine if the distributional assumptions are satisfied and data may be limited. The present paper uses simulation studies to evaluate how well MLE performs by considering the sensitivity of the maximum likelihood estimates to the true parameter values and the distributional assumptions.

A Bayesian Method To Assess The Fit Of N-Mixture Models Used In The Estimation Of Animal Abundance

✦Sherwin Toribio, University of Wisconsin - La Crosse, 1725 State St., La Crosse, WI 54601, *toribio.sher@uwlax.edu*; Jason Rubbert, University of Wisconsin - La Crosse

Key Words: N-Mixture, Abundance Estimation, Bayesian, PPMC, MCMC

Estimating the abundance (or population size) of animals is a major concern within wildlife statistics. One convenient sampling design used to gather information about animal abundance is the simple count method. In this sampling procedure, observers visit randomly selected

sites several times and record the number of animals of a certain species per visit. Although this sampling method is more convenient than the capture-recapture method, the statistical methods needed to obtain the estimates of animal abundance is much more complicated. In 2004, Andrew Royle proposed a statistical method, using N-mixture models, to obtain abundance estimates from this kind of data. However, this method is only good if the model assumptions are satisfied. If some of the assumptions are not true, we have discovered that the estimates from this procedure can be very biased. In this talk, a Bayesian method is presented that can effectively detect violations of some of the model assumptions in the data.

Examining The Bias Of The Lincoln-Petersen Estimator In Two-Sample Closed Population Models

✦Hung-yu Pan, National Chia-Yi University, N0. 300 Syuefu Rd., Department of Appiled Mathematics, Chia-Yi, International 60004 Taiwan, R.O.C., *hypan@mail.ncyu.edu.tw*

Key Words: capture-recapture, cross-product ratio, time effect, behavioral response, heterogeneity, Lincoln-Petersen estimator

We consider the problem of estimating the closed population size for two-sample capture-recapture experiments. For this problem, the Lincoln-Petersen method has been routinely used. However, this approach may result in a biased estimator of the population size for models with some of the major sources of variation in the capture probabilities such as behavioral response and heterogeneity. In this paper, we theoretically examine the bias of the Lincoln-Petersen estimator when the assumptions of equal catchability and independence between samples do not hold. A method to evaluate the magnitude of the bias is developed, based on the cross-product ratio of the capture probabilities in the two samples. The performance of this method is demonstrated through simulation study.

A Bayesian Adaptation Of Publication Bias Compensation For Stranded Marine Mammal Growth Estimation

✦Mary Shotwell, Medical University of South Carolina, , *met106@hotmail.com*; Elizabeth H Slate, Medical University of South Carolina

Key Words: pseudo-data, strandings, publication bias, detection bias

Federal legislation restricting research on wild marine mammals places emphasis on the importance of fully utilizing stranded animals. Bias in detection of stranded marine mammals poses a statistical concern, analogous to publication bias. A Bayesian adaptation of a pseudo-data method for publication bias is described and applied to South Carolina strandings data in order to better estimate total body weight growth of bottlenose dolphins (Tursiops truncatus). A simulation study was performed to assess sensitivity of the method to sample size, detection scenarios, and guesses at the true parameters. Results of the simulation show that the new method outperforms complete case analysis as long as bias is present. The real data analysis results suggest that South Carolina bottlenose dolphins reach smaller sizes at a faster growth rate than previously estimated from complete case analysis.

Modeling Effects Of Climate Change On Pallid Sturgeon In The Missouri River

✦Rima Dey, Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO 65211-6100, *rd7r9@mail.mizzou.edu*; Christopher K. Wikle, University of Missouri; Mark L. Wildhaber, U.S. Geological Survey, Columbia Environmental Research Center; Edward H. Moran, U.S. Geological Survey, Columbia Environmental Research Center ; Christopher J. Anderson, Geological and Atmospheric Sciences, Iowa State University; Kristie J. Franz, Geological and Atmospheric Sciences, Iowa State University

Key Words: Climate change, Hierarchical Bayesian computation, Scaphirhynchus albus, Uncertainty, Bioenergetics, Population model

Pallid sturgeon (Scaphirhynchus albus) are rare in the Missouri river basin and were federally listed as endangered in 1990. To protect this fish population, reliable mathematical models are needed to help predict the response of populations to potential climate change and quantify the predominant sources of uncertainty. This may suggest how to mitigate the adverse effects of climate change on sturgeon populations and could prove useful to identify research questions focused on recovery efforts. Assessing the effect of climate change on the ecosystem requires implementation of multiscale climate models in a hierarchical framework. Climate projections made by global climate models (GCMs) must be downscaled by regional climate models to obtain regional projections that can be linked with watershed models, river hydraulic models and population models in a way that sensibly propagates uncertainty across scales. This talk focuses on implementation of an individual-based bioenergetics model for pallid sturgeon, which provides a means for quantifying the relative importance of various environmental factors on individual growth and consumption given various climate scenarios.

357 Novel Statistical Approaches in Genetic Epidemiology

Section on Statistics in Epidemiology, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Likelihood Ratio Test For Detecting Gene (G)-Environment (E) Interactions Under Additive Risk Models Exploiting G-E Independence For Case-Control Studies

✦Summer S Han, National Cancer Institute, 1255 New Hampshire Ave. NW, APT. 408, Washington, DC 20036 US, *summer.han@aya.yale.edu*; Philip S Rosenberg, National Cancer Institute; Nilanjan Chatterjee, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, USA

Key Words: GxE interaction, Additive risk model, multiplicative risk model, case-control study, gene-environment independence

There has been a long-standing controversy in epidemiology on deciding an appropriate risk scale for testing interactions between gene (G) and environmental exposure (E). Although interaction tests based on multiplicative risk models have been more widely applied due to its convenience in statistical modeling, interactions under additive

risk models have been regarded to be closer to true biological interactions and considered to be more useful in intervention-related decision making processes in public health. It's well known that exploiting the independence information between G and E increases the power of interaction tests. Such approaches, however, have been limitedly applied to the multiplicative scale interaction tests. In this article, we propose a likelihood ratio test for detecting additive scale interactions for case-control studies that incorporates the G-E independence information. Simulation study was conducted to compare the performances of our approach and traditional methods including synergy index (SI), which do not take into account the independence information. We illustrate our method by application to National Cancer Institute lung cancer and smoking data.

Unified Analysis Of Secondary Phenotypes In Case-Control Association Studies

✦Arpita Ghosh, National Cancer Institute, NIH, , *ghosha3@mail.nih.gov*; Fei Zou, The University of North Carolina at Chapel Hill ; Fred Andrew Wright, Univ North Carolina

Key Words: case-control association study, secondary phenotype, secondary analysis, retrospective likelihood, profile likelihood, pseudo likelihood

It has been repeatedly shown that in case-control sampled association studies, analysis of the secondary phenotypes ignoring the sampling scheme can produce highly biased risk estimates. Although a number of approaches have been proposed to properly analyze secondary phenotypes, these approaches often fail to reproduce the marginal logistic model typically assumed for the original case-control phenotype. In addition, handling covariates, especially continuous ones, in a flexible manner remains challenging. We provide a general retrospective likelihood framework to perform association testing for both binary and continuous secondary phenotypes while respecting desired marginal models and allowing for interaction between the genetic variant and the secondary phenotype on primary disease risk. We use a profile likelihood technique to handle the covariates and provide an easy algorithm for deriving the estimator. We also present an alternative approach to handling the covariates based on the pseudo likelihood method. We describe extensive simulations to evaluate the performance of the profile likelihood method in comparison with the pseudo likelihood and other competing methods.

Network-Guided, Sparse Regression Modeling For Detection Of Gene-Gene Interactions

✦Chen Lu, Boston University, Boston University School of Public Health, 801 Mass Avenue, Crosstown Center, 3rd floor, Boston, MA 02118, *chenlu@bu.edu*; JosÈe Dupuis, Boston University; Eric Kolaczyk, Boston University

Key Words: gene-gene interaction, sparse regression, network, pathway, penalized regression

Unlike Mendelian diseases, in which disease phenotypes are largely driven by mutation in a single gene locus, complex diseases are believed to be caused by a number of factors of both genetic and environmental nature, as well as lifestyle. Progress in this area must therefore come from unraveling the interplay of genes with each other, as well as with environment, through the system of biological pathways and related networks. We present a novel, network-guided statistical methodology to facilitate the discovery of gene-gene interactions associated with complex quantitative traits in human disease. Our method uses sparse regression principles to fit regression models (i.e., of phenotype on genetic markers) with second-order interactions, based on a penalized least-squares criterion, where the penalty incorporates known network biology in the form of pathways and gene function. We discuss the derivation of our methodology, as well as its implementation, and present results from simulation studies and various applications.

A Cautionary Note On Testing For Association Of Low-Frequency Variants In Case-Control Studies

✦Guan Xing, Bristol Myers Squibb, Princeton, NJ 08534, *guan.xing@bms.com*; Chao Xing, University of Texas Southwestern Medical Center

Key Words: Wald test, case-control study,, low-frequency variation

Wald test is commonly employed test in genetic association studies because of its computational simplicity. Hauck and Donner (1977) observed an anomalous behavior of Wald test that in a binary logit model, as the distance between the parameter estimate and the null value increases, the test statistic decreases to zero. The human genetic mapping enters an era of whole-genome sequencing aiming at directly identifying disease predisposing variants, the frequency of which is constrained to be low under the pressure of selection. In a sample of case-control study, it is likely that this variant appears only scarcely in cases, but not at all in controls. Thus the anomalous behavior of Wald test raises concerns on using it. In this study, we compared the behavior of four tests: likelihood ratio test, Wald test, score test, and exact test in such a genetic scenario with and without considering multiple covariates. We further compared the methods in a real data of whole-genome nonsynonymous variants screening. To conclude, the Wald test should be used with caution in testing for association of low-frequency variants in case-control studies.

Evaluation Of Removable Statistical Interaction In Cancer Epidemiology Studies

✦Jaya M Satagopan, Memorial Sloan-Kettering Cancer Center, 307, East 63rd Street, New York, NY 10065, *satagopj@mskcc.org*; Robert C Elston, Case Western Reserve University

Key Words: multi-stage carcinogenesis process, analysis of variance, score statistic

Cancer epidemiology studies have traditionally focused on investigating gene-gene and gene-environment interactions as part of the attempt to understand the role of risk factors associated with tumorigenesis or tumor-related traits. This talk focuses on evaluating the omnibus null hypothesis of no removable statistical interaction between two sets of risk factors in cancer epidemiology studies. The statistical interaction between two or more risk factors measures the change in the effect on the outcome of one risk factor when the value(s) of the other risk factor(s) is (are) altered. A statistical interaction is deemed removable if the relationship between the outcome and the risk factors can be made linear through suitable transformations so that the resulting relationship takes the form of a simple additive model. We will show how the multi-stage carcinogenesis paradigm provides insights about testing for removable statistical interactions in a parsimonious manner,

and demonstrate conditions under which this approach can provide a powerful approach to test for interactions relative to an alternative standard approach.

Hierarchical Mixture Of Component Models For Hidden Stratified Data With Applications To Genomic Association Study

✦Yulan Liang, University of Maryland,Baltimore, 655 West Lombard Street, Room 404K,, Baltimore, MD 21201, *ylian001@umaryland.edu*

Key Words: genomic association study, Hierarchical Mixture of Component Models, population stratification, HMM, a generalized mixed linear model, Myocardial Infarction

In this paper, we propose and develop a new class of hierarchical mixture of component model for identifying the heterogeneity of the admixed populations and testing the significance of genetic and environmental factors associated with complex disease while addressing population stratification and confounding issues. This proposed model is able to joint model the genetic map markers information for estimating the population stratification and other environment exposures that are related to complex diseases simultaneously. The high level of model is a mixture model while low level models include HMM model in estimations of individual admixture proportions; and density function of subpopulation with a generalized mixed linear model. This class of models can apply to a wide range of a variety of clinical design scenarios of population-based studies. We evaluate the power and false positive rate of the proposed modeling approaches through Monte Carlo simulation studies and comparison with popular method. Application of the proposed model to empirical population genomic data for myocardial infarction is illustrated. The results and discussions are provided.

Analyses Of High-Throughput Sequencing Data From Cancer Samples

✦Su Yeon Kim, University of California, Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, *skim@stat.berkeley.edu*; Terence Speed, University of California, Berkeley

Key Words: cancer genomics, next-generation sequencing, quality control, somatic alterations, somatic mutations

Analyses of genomic alternations and gene expression levels expedite pinpointing underlying mechanisms for cancer development and developing cancer diagnosis and therapy. Rapid development in high-throughput sequencing technologies allows additional advances due to the un-precedented level of information contained in the sequence data. A vast amount of sequencing data is rapidly being produced and one such example is The Cancer Genome Atlas (TCGA) project. For several tumor types, hundreds of exomes and transcriptomes are sequenced from tumor and the matched normal tissues. Developing efficient tools for processing the data and doing quality control (QC) is essential for improving downstream analyses. In this talk, I will present QC metrics for exome DNA sequencing data, and illustrate how our methods can contribute to reducing false positives in downstream analyses such as identifying somatic mutations.

358 Data Streams, Web Pages & Image Analysis ●

Section on Quality and Productivity

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Non-Stationary Network Traffic Diagnosis Under Correlation Context

✦Yingzhuo Fu, University of California, Riverside, 900 University Avenue, Riverside, CA 92507, *yfu001@ucr.edu*; Daniel R. Jeske, University of California, Riverside

Key Words: non-stationary, correlated data, GLMM

Streams of network data are usually correlated, both in short time periods and also in longer time periods. Network data also exhibits non-stationary in the mean structure, frequently with discernable diurnal patterns for example. Data structures of this type present challenges when trying to use standard change-point detection algorithm. We employ a Generalized Linear Mixed Model (GLMM) to model the correlation with embedded random effects, and we capture non-stationarity in the mean structure with fixed effects. The GLMM paradigm also allows the response variable , which is often counts, to be modeled directly. The key step in our modeling process is building the baseline GLMM using historical cycles of the data streams. Conditional on the predicted future realization of the random effects, the data are independent and potential sample paths for future data can be simulated. A transformed CUSUM tracking statistics is then used to detect changes. A control limit is chosen according to a specific false alarm rate. We will illustrate the use of our proposed model with data collected from a real net work and conduct a simulation study to characterize its ability to detect changes.

Nonparametric Sequential Change-Point Procedure For Network Surveillance Data

✦Tatev Ambartsoumian, Department of Statistics, University of California, Riverside, 2981 Elgin Dr, apt A, Riverside, CA 92507 USA, *tamba001@ucr.edu*

Key Words: nonparametric, change point, cusum, generalized likelihood ratio test, data monitoring

We adapt the Generalized Likelihood Ratio Test based Cusum procedure to detect a change in the distribution of the incoming sequence in the context of computer network monitoring. For the case of the known in-control and out-of-control densities, we propose an analytical approximation formula for determining the threshold value of the GLRT Cusum. We further extend the application of the GLRT Cusum for the case of unknown densities under the following two assumptions: 1) there is enough historical data to estimate the in-control distribution, and 2) the out-of-control density can be derived from the in-control density through a proper transformation. We adjust our analytical approximation formula for the GLRT Cusum threshold so that it can be applied under the unknown densities scenario as well. The use of our proposed nonparametric GLRT Cusum technique and the threshold approximation formulas is illustrated by means of several examples.

Website Monitoring And Improvement

✦Roger Longbotham, Microsoft, One Microsoft Way, Redmond, WA 98052, *Roger.Longbotham@microsoft.com*; Ji Chen, Microsoft Corporation; Dave DeBarr, Microsoft Corporation; Shaojie Deng, Microsoft; Justin Wang, Microsoft Corporation

Key Words: website, monitoring, analytics, experimentation, metrics, content optimization

There are specific statistical issues and methodologies for monitoring and improvement of websites. We will discuss metrics, limitations and pitfalls to ongoing website monitoring and experimentation to improve website performance. The authors have many years of experience with a number of websites and will share best practices and lessons learned. Specific topics include online experimentation principles, feedback analysis and monitoring, visitor segmentation and content optimization.

Data Quality For Online Experimentation

✦Ji Chen, Microsoft Corporation, One Microsoft Way, Redmond, WA , *jich@microsoft.com*; Roger Longbotham, Microsoft; Justin Wang, Microsoft Corporation; Shaojie Deng, Microsoft; Dave DeBarr, Microsoft Corporation

Key Words: online experimentation, data quality, web analytics, anomaly detection

Controlled experimentation has been proven to be an effective way to test ideas and evaluate changes in websites and web services. While the basic theoretical foundation for controlled experiments has been well established, in reality more often than not, we are faced with data quality issues that could easily bias the results of the experiments and confound the decision making process. This talk will discuss several data quality concerns specific to online experimentation and provide best practices to address them. We will cover challenges such as web robot detection, traffic anomaly alerts, user session identification, page instrumentation issues and web data cleansing. Most of the techniques discussed are also applicable to web analytics in general. Some research questions will be presented at the end.

Framework For Measurement And Prevention Of Human Error In Service Delivery

✦LARISA SHWARTZ, T.J.Watson Research, IBM, 19 Skyline Dr, Hawthorne, NY 10532 USA, *lshwart@us.ibm.com*; Genady Grabarnik, St. John's University

Key Words: human error, service delivery, automation and process improvement

The theory of human error addresses interactions between human performance variability and situational constraints. The occurrence and frequency of human error more greatly depend on interactions with environment than stable and inherent characteristics of the operator or task. In our study, we applied two approaches to understanding the pervasiveness of human error. We analyzed incident tickets from a large service provider and surveyed service delivery experts (SMEs). We requested that the SMEs classify incidents into three categories: 'human error,' 'not a human error,' or 'maybe caused by human'. We discovered a significant discrepancy in the estimation we got from these approaches. We examined and evaluated the reasons for the discrepancy and concluded that quantitative measures, such as error rates, could be inadequate and misleading. Thus, meticulous and consistent descriptions of the error-prone conditions must be considered. Furthermore, we considered an impact of automation on human error prevention and process improvement, and provided recommendations for the usage of these methods, illustrating them on specific use-cases.

A Spatiotemporal Method For The Monitoring Of Image Data

✦Fadel M. Megahed, Virginia Tech, 250 Durham Hall, Blacksburg, VA 24061, *fmegahed@vt.edu*; Lee J. Wells, Virginia Tech; Jaime A. Camelio, Virginia Tech; William H. Woodall, Virginia Tech

Key Words: Change-point Model, High Density Data, Image-based Monitoring, Profile Monitoring, Statistical Process Control, Steady State

Machine vision systems are increasingly being used in industrial applications due to their ability to provide information on product geometry, surface defects, and/or surface finish. Previous research for monitoring these visual characteristics using image data has focused on either detecting changes within the image or between images. Extending these methods to include both the spatial and temporal aspects of image data will provide more detailed diagnostic information. Therefore, in this paper, we show how image data can be monitored using a spatiotemporal framework that is based on the use of a generalized likelihood ratio control chart. The performance of the proposed chart is evaluated with respect to an image-based profile monitoring technique through simulations and experimental work. The results show that our GLR chart outperforms the profile monitoring method. More importantly, the simulations show that our method provides a good estimate of the change-point and the size/location of the fault, which are important fault diagnostics' metrics that are not typically reported. Finally, we highlight some research opportunities and provide some advice to practitioners.

Use Of Image Analysis Methods In Nondestructive Evaluation

✦Ye Tian, Iowa State University, 1414 Snedecor Hall, Ames, IA 50011, *tianye@iastate.edu*; William Q. Meeker, Iowa State University; Ranjan Maitra, Iowa State University

Key Words: Thermography, Ultrasonics, Matched filter, Probability of detection, Signal-to-noise ratio (SNR)

Traditional nondestructive evaluation has been done by taking a signal response (e.g., a voltage) and mapping it into a scalar that can be used for detection decision making. Modern nondestructive evaluation techniques for flaw (e.g., a crack) detection (e.g., ultrasonics, X-ray, and thermography) have images for outputs. There is a need to develop automatic crack detection algorithms to reduce the substantial effects of human-factors variability in making crack-detection decisions. In this paper, we develop and compare several methods for mapping an image into a statistically efficient decision criterion. For example, the output of a matched filter can be evaluated with a signal-to-noise ratio (SNR) criterion. After obtaining the SNR related metrics through dynamic feature extraction, we implement a statistical algorithm to automatically detect the crack's existence or not. Similar approach can be used with other kinds of filtering. We use probability of detection to compare the different approaches.

359 Contributed Oral Poster Presentations: Biometrics Section

Biometrics Section

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Multivariate Random Effects Meta-Analysis: An Extension

✦Kepher H Makambi, Georgetown University, 3800 Reservoir Road, NW, Lombardi Comprehensive Cancer Center, LL S176, Washington, DC 20057, *khm33@georgetown.edu*; Hyunuk Seung, Georgetown University

Key Words: Heterogeneity variance estimator, Weigting, Bias, Mean square error

In the random effects univariate and multivariate meta-analysis, a number of iterative and non-iterative approaches have been proposed including the DerSimonian and Laird (D-L), maximum likelihood (ML) and restricted maximum likelihood methods (REML). In this study, we propose a multivariate approach based on procedures developed by Hartung and Makambi (2002); with arguments along the lines of Jackson et al. (2009). A simulation study demonstrates that our approach is competitive compared with the commonly used approaches in the extant literature. We present examples to illustrate the application of our procedure.

Identifying Single Feature Polymorphisms Using Affymetrix Gene Expression Data

✦Cumhur Yusuf Demirkale, University of Maryland, Baltimore, MSTF 261 A, 10 Pine Street, Baltimore, MD 21201, *cyusufa@gmail.com*

Key Words: Affymetrix Gene Chips, Linear Mixed Models, Microarrays, Outliers, Robust Methods, Single Feature Polymorphisms

In microarray data analysis, the identification of Single Feature Polymorphisms (SFPs) is important for producing more accurate expression measurements when comparing samples of different genotypes. Also, portions of DNA that differ between parental lines can serve as markers for tracking DNA inheritance in offspring. We summarize several SFPs discovery methods in the literature. To identify single probe defining SFPs in the data, we developed two new algorithms where a difference value is defined for each probe after accounting for the overall gene expression level differences in the probe set. First method contrasts the difference value of each probe with the average of the difference values for the rest of the probes in that probe set. Second method is a robust version of the first method. The performances of all methods are compared through two publicly available published data sets, where truth about the sequence polymorphism is known for some "Gold Standard" probes.

R Library To Validate A Gene Signature In Multiple Gene Expression Datasets, Using Principle Component Analysis

✦William James Fulp, Moffitt Cancer Center, MRC-BIOSTAT Room 2062, 12902 Magnolia Drive, Tampa, FL 33612-9416, *William.Fulp@Moffitt.org*; Dung-Tsa Chen, Moffitt Cancer Center

Key Words: Gene Signature Validation, Microarray, Principle Components Analysis, R Library

Validation of a gene signature often requires its test on independent datasets to demonstrate the association of the signature with clinical outcomes. While many public gene datasets exist, cleaning these datasets and validating a gene signature is nontrival. To quickly allow for gene signature validation, we use R software to develop libraries to collect gene expression data for a specific cancer type, and an algorithm based on principal component analysis to generate a weighted summary score to reflect the combined effect. Moreover, the library includes functions to generate various plots, such as survival curves or boxplots, depending on clinical outcome, for the weighted summary score and univariate analysis results. To date, we have created two libraries, one for breast cancer (12 datasets), and one for ovarian cancer (6 datasets). By utilizing our R libraries for gene expression data, we can test multiple new gene signatures for their clinical association in a timely fashion. We are working on building libraries for other cancer types, such as colon cancer and lung cancer.

A Two-Stage Linear Mixed/Cox Model for Longitudinal Data and Disease Progression in Pancreatic Cancer Patients

✦Wei Qiao, MD Anderson Cancer Center, 1400 Pressler Street, Houston, 77030, *wqiao@mdanderson.org*; Ning Jing, The University of Texas Health Science Center at Houston

Key Words: two-stage model, linear mixed model, longitudinal, Cox model

Cancer studies often collect time-to-event data and repeated measurements, e.g. biomarkers, for each subject. In many cases, the research interest is to evaluate if the individual level and progression rates of repeatedly measured biomarkers can quantify the severity of the disease and predict the subject's susceptibility to disease progression. In this study, we present a two-stage model that takes into account the dependency and association between longitudinal measurements of biomarkers and time-to-event data. In the first stage, the subject-specific biomarker trajectories need to be modeled and estimated using linear mixed-effect model ; in the second stage, the subject-specific biomarker trajectories estimated from the first stage are used as coviarates in the Cox model for the disease progression. Consequently, the effects of the biomarker trajectories on the time-to-progression can be assessed. The information matrix of the partial likelihood in the second stage cannot be used to make inference for the estimated risk coefficients, since it does not take into account the uncertainty of the estimated biomarker trajectories. Alternatively, the bootstrap procedure can be u

Dirichlet-Multinomial Power Calculations And Statistical Tests For Microbiome Data

✦Patricio La Rosa, Washington University School of Medicine, 660 S Euclid Ave, Box 8005, St Louis, MO 63110, *plarosa@dom.wustl.edu*; William Shannon, Washington University School of Medicine; Elena Deych, Washington University School of Medicine; George Weinstock, Washington University School of Medicine; Erica Sodergren, Washington University School of Medicine; Paul Brooks, Virginia Commonwealth University; Edward Boone, Virginia Commonwealth University; Qin Wang, Virginia Commonwealth University

Key Words: Microbiome, dirichlet-multinomial, next-gen sequencing, power/sample size, diversity indices, applied biostatistics

We provide a statistical framework to perform formal hypothesis testing, and to calculate power and sample size requirements for human microbiome experiments using taxonomical classification of metagenomic sequences. The methods proposed allow for modeling and comparing statistically the taxa abundance distributions of microbiotas from one and several populations. In particular, we use the Dirichlet-Multinomial (DM) distribution to model taxa counts from a set of samples. The DM model takes into account the variability of the taxa probabilities or relative abundance distribution (RAD) across samples providing a better fit to the data than a Multinomial model. We study the power and size of the following hypothesis tests: Multinomial goodness of fit against a Dirichlet multinomial alternative; one-sample, two-sample, and multiple-sample comparison of RAD means; and two-sample comparison of RAD distributions. More specifically, for a given taxa number, we provide guidelines for computing sample size, namely, the numbers of subjects and number of reads per subject required to obtain a desired statistical power. As an example, we apply our methodology to the HMP data on 24 subjects.

Statistical Analysis Of Taxonomic Trees In Microbiome Research

✦William Shannon, Washington University School of Medicine, 660 S Euclid Ave, Box 8005, St Louis, MO 63110, *wshannon@wustl.edu*; Patricio La Rosa, Washington University School of Medicine; Elena Deych, Washington University School of Medicine; Yanjiao Zhou, Washington University School of Medicine; George Weinstock, Washington University School of Medicine; Erica Sodergren, Washington University School of Medicine; Berkley Shands, Washington University School of Medicine

Key Words: Object data analysis, taxonomic trees, microbiome, genetic sequencing, applied biostatistics

Human microbiome research uses next generation sequencing to characterize the microbial content from human samples to begin to learn how interactions between bacteria and their human host might impact health. As an emerging medical research area there are few formal methods for designing and analyzing these experiments, with most approaches being ad hoc and applicable to the particular problem being faced. Since microbiome samples can be represented as taxonomic trees, it is natural to consider statistical methods which operate on graphical structures such as tree objects. A unimodal probability model for graph-valued random objects has been derived and applied to several types of graphs (cluster trees, digraphs, and classification trees). In this work we apply this model to HMP taxonomic trees which allows for a fully statistical data analysis. This model allows us to calculate core microbiomes using statistical maximum likelihood estimation, test hypotheses and calculate P values of whether the core microbiomes are the same or different across patient subgroups using likelihood ratio tests. As an example, we apply our methodology to the HMP data on 24 subjects.

A New Canonical Correlation Association Measure For Multivariate Analysis With Massive Categorical Data

✦Hongyan Chen, Stony Brook University, Chapin L2172D, 700 Health Sciences Drive, Stony Brook, NY 11790, *chenxuan333@hotmail.com*; Laura Jean Bierut, Washington University in St. Louis; Wei Zhu, State University of New York at Stony Brook

Key Words: GWAS, SNPs, categorical, canonical correlation, association measure

Fueled by the modern genome-wide association studies (GWAS), there is a growing need for novel multivariate analysis methods for massive categorical data. The goal of GWAS is to investigate the relationship between phenotypes and genotypes, which can be determined by over a million single nucleotide polymorphisms (SNPs). Such high dimensionality demands more efficient multivariate analysis tools such as cluster analysis (CA) and partial correlation network analysis (PCNA). However, both of them are mainly developed for the numeric data. An urgent task is to customize them for categorical one by developing suitable association/distance measure. In this paper, we first examine the performance of several existing measures between SNPs. Subsequently, by treating SNPs as categorical variables, we propose a novel (partial) canonical correlation association measure that has significant advantages. In applying this new association measure to COGEND, a small GWAS study on nicotine-addiction, We discovered that the CA achieved more accurate chromosomal separation and the hub-SNPs identified by the PCNA yielded higher prediction accuracy for nicotine-addiction status.

Coupling A Frequency-Based Feature Selection Method And Voting Classifier Approach For Biomarker Detection

Sandra L Taylor, University of California Davis; ✦Kyoungmi Kim, University of California Davis, Division of Biostatistics, One Shields Ave, Davis, CA 95616 USA, *kmkim@ucdavis.edu*

Key Words: Biomarker detection, feature selection, classification, omics

With technological advances, the -omics fields have recently explored and applied to develop diagnostic and prognostic tests. An investigator's objective is to develop a classification rule to predict class of unknown samples based on a small set of features that can ultimately be used as biomarkers in a clinical setting. A number of methods have been developed for feature selection and classification using gene expression data, proteomics or metabolomics data. However, the great challenge is not the design of algorithms but the potential application in a clinical setting. While common classification methods such as random forest (RF) and support vector machines (SVM) are effective at separating groups, they do not directly translate into a clinically-applicable classification rule based on a small number of features. In this study, we present a frequency-based feature selection method coupled with a voting classifier approach that yields "stable" feature sets. We evaluate the performance of voting classifiers using three -omics datasets. We show our approach achieves classification accuracy comparable to RF and SVM while yielding classifiers with clear clinical applicability.

A Modified Pseudo Maximum Likelihood Estimate For Roc Curve Of A Generalized Odds-Rate Model

✦Huining Kang, University of New Mexico, HSC DOIM, MSC10 5550, 1 University of New Mexico, Albuquerque, NM 87131-0001, *HuKang@salud.unm.edu*; Edward Bedrick, University of New Mexico Health Sciences Center

Key Words: ROC curve, Proportional odds-rate model, Pseudo maximum likelihood estimator

Previously we proposed a new functional form for the receiver operating characteristic (ROC) curve derived from the generalized proportional odds-rate (GPO) model and a pseudo maximum likelihood estimator (PMLE) for it. In this presentation we propose a modified pseudo maximum likelihood estimator (MPMLE) and provide the asymptotic distribution theory and procedure for using it to make statistical inference for the parameters of GPO ROC curves. The validity of the asymptotic distribution for making inference in finite sample is verified by simulation studies. We also compared the relative efficiencies of the two estimators. As a result, MPMLE has superior efficiency over PMLE. Finally we note that the two estimators can be used to estimate the ROC curve with an arbitrary parametric form.

Determinants Of Fecal Bile Acids

✦Weiqun Tong, University of Arizona, , *weiqunt@email.arizona.edu*; Paul Hsu, University of Arizona; Patricia Thompson, University of Arizona

Key Words: Bile Acids, colorectal cancer, 2nd bile acid, Two-stage analysis

Colonic bile acids exposure, especially secondary (2nd) bile acids, is associated with higher incidence of colorectal cancer (CRC). We conducted a secondary, cross-sectional analysis for 735 participants in a phase III clinical trial of Ursodeoxycholic acid for the prevention of colorectal adenoma. Different categories of baseline fecal acid levels were selected as outcome variables. Dietary, non-dietary, medication use, body mass index, and physical activity were evaluated as determinants of fecal bile acids. A two-stage analysis (logistic regression followed by linear regression) was applied if the detection of bile acid was lower than 50%; and for those 50% or more were detectable, a multiple linear regression model was used. Low density lipoprotein (LDL), triglyceride, fiber intake, and use of statin drugs were associated with total 2nd (P = 0.016, < .001, 0.001, 0.003 respectively) and total fecal bile acid levels (P= 0.011, < .001, 0.001, 0.004 respectively). Being male was associated with 25.5% lower total 2nd bile acid levels. Among the factors that we assessed, our results indicate that blood lipids are the strongest determinants of fecal bile acids levels.

An Empirical Evaluation Of Array Normalization For Agilent Microrna Expression Arrays

✦Li-Xuan Qin, Memorial Sloan-Kettering Cancer Center, 307 E 63rd St, New York, NY 10065, *qinl@mskcc.org*; Jaya M Satagopan, Memorial Sloan-Kettering Cancer Center; Sam Singer, Memorial Sloan-Kettering Cancer Center

Key Words: microarray, microRNA, normalization

Methods for array normalization have been developed for mRNA expression arrays, such as median normalization and quantile normalization. These methods assume few or symmetric differential expression of markers on the array. The performance of the existing normalization methods need to be re-evaluated when applied to microRNA arrays, which consist of a few hundred markers and a reasonable fraction of them are anticipated to have disease-relevance. We empirically examined sources of variations in miRNA array data using a set of Agilent arrays in liposarcoma (n=56) and evaluated normalization methods using Solexa sequence data on a subset of these tumors (n=29) as the gold standard. We found that there is minimum variation between replicate probes for the same target sequence and moderate variation between multiple target sequences for the same miRNA. There is moderately high correlation between Agilent data and Solexa data. Quantile normalization has slightly improved the correlation with Solexa data, as well as the detection of differentially expressed microRNAs both in terms of statistical significance and the direction of change.

Mixed Models In The Genetic Mapping Of The Cardiovascular Risk Factors In Brazilian Families Using Snps Data

✦Mirian Souza, Instituto de Matematica e Estatistica da Universidade de Sao Paulo Brazil, Rua Lucindo Passos Filho, 127 Jd Almanara, Sao Paulo, International 02865040 Brazil, *miria_sou@hotmail.com*

Key Words: complex diseases, mixed models, family data

The study of complex diseases such as hypertension, diabetes and obesity is of great importance in medicine since these diseases affect many persons in our country and the world. It is believed that the pattern of variation of these diseases involves environmental and genetics components, and their possible interactions. For mapping of genes in human complex diseases, the biotechnology resources have increased very fast allowing that more dense markers maps are available. However, the development and use of analytical methodologies to analyze genetic data set have not followed on the same velocity. In this context, SNPs (Single Nucleotide Polimorphisms) platforms consisting of a million of loci represent a big analytic challenge. In this work, for mapping of genes associated with cardiovascular risk factors will be considered different mixed models formulations in the analysis of family data and SNPs platforms. To implement the proposed methodologies will be investigated computational facilities with software R and application of these mixed models in real or literature data. Also there intend to discuss the limitations and advantages of the presented mixed models.

Evaluation Of Time To Event Data In The Presence Of Informative Censorship

✦Susan Y Zhou, CDER/FDA, 10903 New Hampshire Ave., Silver Spring, MD 20993, *susan.zhou@fda.hhs.gov*; Guoxing Soon, CDER/FDA

Key Words: Kaplan Meier, Hodge Lehmann Estimator, informative censorship, imputation, simulation

When analyzing time-to-event data for the estimation of treatment effect, we noticed significant discrepancies using different approaches such as Kaplan Meier (KM) and Hodge Lehmann Estimator (HLE). In the clinical trial setting it may not be as evident that the censoring is non-informative. Hence the use of the KM may not be suitable.

To resolve this issue, several imputation scenarios including worst-case scenario were conducted to complete the missing data and the HLEs were obtained and compared. A simulation study was carried out to further investigate the effects of different patterns of informative censorship associated with the misuse of the KM. Several imputation approaches were performed on missing data prior to the estimation of treatment differences. In the present work, the median of all paired treatment difference placebo vs. treatment was obtained for the HLE, and the difference in median time between placebo arm and treatment arm was used for the KM approach. We will summarize the performances of different statistical approaches for selected missing patterns and discuss alternative analysis approaches for time-to-event data with informative censoring.

Distribution-Based Normalization To Reduce Noisy Variability In Small Sample Size Microarray

✦Hui Xie, TRI-Florida Hospital, 2566 Lee Road, Winter Park, FL 32789, *hui.xie@flhosp.org*

Key Words: Normalization, Microarray, Distribution-based, Small Sample Size, Data augmentation

Background: The quantile normalization method is widely used to compensate for technical bias in microarray data. But when the experimental sample size is small, this shoehorning method (1) inappropriately reduces biological differences due to forcing the extremely high (or low) intensity into same distribution shape (2) distorts the distribution of original data. This, in turn, leads to the extra type II error in the gene differential analysis. Results: Original expression intensity can be split into two parts: true signal & noisy signal. We proposed a distribution-base method to model the noisy signals which can increase normalization accuracy, especially for microarray study with small sample size design. The noisy signal had a Gamma distribution but depended on true signal. The parameters of the distribution were iteratively estimated through data argumentation procedure. Both simulated and real data were used to evaluate the normalization accuracy. The new distribution -based method can be applied to normalize rt-PCR expression data as well.

Impact Of Regression Methods On Detecting Increased Coronary Artery Calcification In Rheumatological Diseases

✦Tebeb Gebretsadik, Vanderbilt University School of Medicine, , *tebeb.gebretsadik@vanderbilt.edu*; Ayumi Shintani, Vanderbilt University

Key Words: skewed response, cluster zeros, regression method, Monte Carlo simulation, coronary artery calcification

The highly skewed distribution of an outcome variable with a large cluster of zeros has led to several analytic challenges. Our motivating example is the study of coronary artery calcification (CAC) among systemic lupus erythematous (SLE) and rheumatoid arthritis (RA) for markers of premature coronary disease. Many analytical methods have been used to analyze highly skewed Agatson score, a measure of CAC where the majority of patients have zero values along with few patients with high extreme values. We conducted Monte-Carlo simulation study to assess the performance of commonly used regression methods including linear regression with transformation (CAC+1), logistic regression (no CAC vs. any), the proportional odds logistic regression, the quantile regression, zero-inflated negative binomial regression and hurdle regression methods. Power and type I error rates were evaluated.

Prediction-Based Structured Variable Selection Through the Receiver Operating Characteristic Curves

Huaihou Chen, Columbia University; ✦Yuanjia Wang, Columbia University, 722 West 168 Street, R-6, Department of Biotatistics, Columbia University, New York, NY 10032 USA, *yuanjia.wang@columbia.edu*

Key Words: Support vector machine, Area under the curve, Disease screening, Hierarchical variable selection

In many clinical settings, a common problem is to assess accuracy of a screening test for early detection of a disease. An example is a study conducted to design a new screening test by selecting variables from an existing screener with a hierarchical structure among variables: there are several root questions followed by their stem questions. It is unreasonable to select a model that only contains stem variables but not its root variable. In this work, we propose methods to perform variable selection with structured variables when predictive accuracy of a diagnostic test is the main concern of the analysis. We take a linear combination of individual variables to form a combined test. We then maximize a direct summary measure of the predictive performance of the test, the area under a receiver operating characteristic curve, subject to a penalty function to control for overfitting. We cast the problem of maximizing predictive performance of a combined test as a penalized support vector machine problem and apply a re-parametrization to impose the hierarchical structure among variables. We apply developed methods to design a structured screener to be used in primary care clinics.

Sharing Information Across Genes To Estimate Overdispersion In Rna-Seq Data

✦Steven Peder Lund, Iowa State University, Department of Statistics, Snedecor Hall, Ames, IA 50011, *lunds@iastate.edu*; Dan Nettleton, Iowa State University

Key Words: RNA-seq data, Generalized Linear Models, Overdispersion

Next Generation Sequencing technology can be used to measure gene expression (mRNA) levels. The resulting RNA-seq datasets consist of integer counts typically ranging from 0 to several thousand, with many observations less than 5. The low-count integers in RNA-seq datasets suggest using generalized linear models to identify differentially expressed (DE) genes. However, resulting GLM fits often provide evidence of overdispersion in many genes. While many methods exist for estimating overdispersion for a single gene, there are often few degrees of freedom available for these estimates. Here we present a new method for borrowing information across genes to obtain improved estimates of overdispersion parameters. Using these improved overdispersion estimates produces tests that more accurately control type I error rates and improve detection of DE genes.

Ensemble-Based Selective-Voting Algorithm For Cancer Classification

✦Chuanlei Zhang, University of Arkansas for Medical Sciences, , *czhang@uams.edu*; Radhakrishnan Nagarajan, University of Arkansas for Medical Sciences; Eric Siegel, University of Arkansas for Medical Sciences; Ralph Kodell, University of Arkansas for Medical Sciences

Key Words: Classification, Genomics, Cancer, Selective voting, Ensemble, Convex hull

There is much interest in using gene expression data to classify cancer patients. However, sufficient predictive ability for clinical application has not been shown using traditional classification algorithms. Recently, a model-free ensemble algorithm has been proposed for classifying patients using high-dimensional genomic data. In this paper, the convex-hull structure of that algorithm is used to develop a new ensemble-based selective-voting algorithm. This new algorithm allows members of the ensemble to vote when test points fall inside reduced two-dimensional non-overlapping convex hulls defined by pairs of predictor variables. We study two different pruning methods to trim overlapping convex hulls to achieve separation of classes and we investigate various numbers of bi-variate regression models to select gene pairs as predictor variables. Only gene pairs for which either member does not appear in a higher-ranked pair based on the regression R-square are kept as unique sets of potential voters. The algorithm is tested on a publicly available colon cancer dataset. Classification accuracy is shown to be improved using the new algorithm.

Splitting Random Forest Algorithm And Partial Least Square Method In Finding An Pptimal Sets Of Genes For Glioblastoma (Gbm)

✦Xiaowei Guan, Department of Epidemiology & Biostatistics, Case Western Reserve University SOM, 10900 Euclid Avenue., clelveland, OH 44106, *xxg39@case.edu*; Mark Chance, Center for Proteomics and Bioinformatics,Case Western Reserve University, SOM; Abdus Sattar, Case Western Reserve University; Jill Barnholtz-Sloan, Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine

Key Words: Glioblastoma, Splitting Random Forest, Tree-based modeling, Partial Least Square, Likelihood-based method, Gene expression

Glioblastoma multiforme (GBM) is the most common and fatal brain tumor in adults. There is an urgent need to find the most compact set of genes that characterize the different subtypes of GBM so that this set could be easily implemented for clinical care. Standard approaches for analysis of gene expression data include t-tests, analysis of variance (ANOVA) and clustering. In this study we are proposing two new approaches in determining a best set of genes: 1) a splitting random forest algorithm (SRF) based on iteratively splitting a dataset equally into training and testing sets using different running times in order to extract the optimal classification-based gene signature utilizing a random forest bootstrap selection process, and 2) a partial least squares embedded in likelihood based variable selection method, utilizing inverse probability weighting. Necessary simulation will be performed and, we will apply these methods to the publicly available GBM Cancer Genome Atlas gene expression data.

Optimal Power Transformations For Tma Biomarker Data

✦Bhupendra Rawal, Moffitt Cancer Center and Research Institute, 12902 Magnolia Dr, MRC/Biostat, Tampa, FL 33612, *bhupendra.rawal@moffitt.org*; Daohai Yu, Moffitt Cancer Center and Research Institute; Michael J. Schell, Moffitt Cancer Center and Research Institute

Key Words: Transformation, Biomarker

Pre-processing data using raw (R), log (L), square-root (S), or quarter-root (Q) as an optimal transformation has been accepted in observational and clinical studies. There is a special need to look for an optimal transformation in tissue-microarray (TMA) data. 53 TMA runs of biomarker expression from 187 early stage non-small cell lung cancer patients were studied to assess the optimal power transformations. The automated-quantitative analysis (AQUA) was used to measure the nucleus and cytoplasmic scores. The goal of optimal transformation is to homogenize the variability of AQUA scores in order to minimize the influence of any individual score(s). We used the MM robust regression method for estimation and selected as the optimal transformation the one with the lowest average number of high leverage points. 7 (13%) runs favor R for analysis and the other 46 (87%) runs require a transformation. Among the 46 runs, 13 (25%), 13 (25%), and 10 (19%) favor L, S and Q, respectively, while 8 (15%) runs select either L or Q, and 2 (4%) runs either S or Q. Thus, pre-processing biomarker data to determine an optimal transformation is an important step in the statistical analysis of TMA data.

Estimation Of High-Parent Heterosis In Gene Expression

✦Tieming Ji, Department of Statistics, Iowa State University, 27C Schilletter Village, Ames, IA 50011 US, *tji@iastate.edu*; Peng Liu, Iowa State University; Dan Nettleton, Iowa State University

Key Words: empirical Bayes, hierarchical Modeling, hybrid vigor, microarray, statistical genetics

High-parent heterosis, also known as hybrid vigor, occurs when the mean trait value of offspring is greater than the mean trait value of each parent. For maize, this phenomenon was first documented in the late 1800s and is the basis of the seed corn industry today. In an effort to understand the molecular genetic mechanisms responsible for heterosis, researchers have begun to measure the expression levels of thousands of genes in parental maize lines and their offspring. We will present an empirical Bayes approach for estimating the offspring mean minus the maximum of the two parental means (high-parent heterosis) for each of thousands of gene expression levels. We will demonstrate through simulation that our estimators have both lower bias and lower mean square error than naÔve estimators defined by gene-specific differences between the offspring sample mean and the larger of the two parental sample means. We will illustrate the point and interval estimates produced by our method via application to a maize microarray dataset.

Cpain: The Challenges And Pitfalls Of Analyzing Patient Reported Data For Pain Management Therapy

✦Dawn Harwood, REGISTRAT MAPI, 2343 Alexandria Drive, Suite 400, Lexington, KY 40503, *dharwood@registratmapi.com*

Key Words: Patient Reported Data, Pain Management

Chronic pain (defined as intractable pain which has occurred in excess of three months) affects approximately 33% of individuals in the United States. Chronic Pain Impact Network (CPAIN) includes a longitudinal registry sponsored by REGISTRAT MAPI. The registry aims to establish a robust clinical database that will aid in answering questions related to real life management of chronic pain. This database will be analyzed to evaluate the effectiveness of pain management therapy; specific medical events of interest; substance use and medication misuse, abuse and diversion and evaluate the economic impact of pain. In addition, analyses will consider demographics, pain diagnoses and treatments for chronic pain patients and describe the training and practice settings of clinicians managing chronic pain. The subjective nature of the patient reported data makes the statistical analysis challenging. Therefore it is imperative that the data is understood and analyzed in a meaningful way. Data will be presented to highlight these challenges and guide future directions for pain management therapy.

Multiple Analytical Approaches In The Presence Of Missing Data: A Case Study Of Esperanza Y Vida Cancer Screening Intervention

✦Jessica Hersh, University of Arkansas for Medical Sciences, , *jessternick@hotmail.com*; Zoran Bursac, University of Arkansas for Medical Sciences; D. Keith Williams, University of Arkansas for Medical Sciences; Linda Thelemaque, Mount Sinai School of Medicine; Lina Jandorf, Mount Sinai School of Medicine; Deborah O Erwin, Roswell Park Cancer Institute

Key Words: Missing data, Imputation, Logistic regression

Esperanza y Vida (EyV) is a three-site, community based intervention, designed to increase cancer screening rates among Latinas. Three outcomes, including papa test, clinical breast exam and mammogram screening compliance are evaluated at the two month follow-up, among previously non-adherent women. Overall 78% of the participating women had complete follow-up data. Three analytical approaches are compared to determine the treatment effect significance (efficacy) with respect to each of the three outcomes. First one is complete data analysis (CDA), second one assumes non-compliance among those with missing follow-up data (ZCF), and the third one involves logistic regression model based, multiple imputation of the outcome(s) (LMI). All analyses except univariate ZCF, indicate non significant treatment effect. Univariate ZCF is likely significant due to the fact that those with missing follow-up differ in some characteristics as compared to those with complete follow-up data. After the multivariate adjustments however, the treatment effect is non-significant. Implications of these findings are further discussed.

Clustering Analysis In Gene Co-Expression Network

✦ZILU ZHANG, UCLA, 10916 Ashton Ave, Apt 104, Los Angeles, CA 90024, *zhangzilu83@gmail.com*

Key Words: network, clustering, hierarchical, adjacency

While some adjacency matrices are not factorizable, we can define them as approximate conformity-based network. Generalize the relationship between approximate CF-based network to fundemental network concepts. Develope whether the relationship still hold for networks with relatively low factorizability. How to apply hierarchical clustering analysis in gene co-expression network.

Statistical Inference Of Neuronal Functional Connectivity From Parallel Spike Trains

✦Zhanwu Liu, Carnegie Mellon University Department of Statistics, 5000 Forbes Ave, Baker Hall 132, Pittsburgh, PA 15213 USA, *zhanwul@andrew.cmu.edu*

Key Words: functional connectivity, spike trains, computational neuroscience, sample size

In neuroscience study, it is desirable to understand how the neuronal activities are associated and how the association changes with task based on parallel spike train recordings from multielectrode array. The term functional connectivity is used to describe the association between neurons and the change of association with task purpose. One basic question in functional connectivity inference is how much sample size is needed for reliable estimation. In this abstract, different combinations of coupling strength, coupling function and sample sizes were simulated and studied. The difficulties of functional connectivity inference will be discussed.

Use Of Selective Phenotyping To Increase Power Of Genetic Association Studies Of Quantitative Biomarkers

✦Yunfei Wang, University of North Carolina, 5000D,120 Manson Farm Road, Chapel Hill, NC 27599, *wang484@med.unc.edu*; Ethan M. Lange, University of North Carolina

Key Words: biomarkers, simulated annealing, statistical power, genetic association, SNPs

Blood-based biomarkers are often used as intermediate outcomes for identifying genetic risk factors associated with cardiovascular disease (CVD). Measuring biomarkers, however, is typically expensive and time consuming. Genome-wide genetic data on single-nucleotide polymorphisms (SNPs) are now routinely available for tens of thousands of samples from large population-based cohorts. Given the expense of measuring biomarkers, it would be desirable to identify a subset of subjects that could be phenotyped for the biomarker of interest in order to optimize statistical power under fixed cost constraints. For any specific SNP and a fixed sample size, power is typically optimized when genotypes are partitioned equally between homozygotes for the major and minor allele. When trying to optimize power across multiple SNPs, using a selection strategy that optimizes power for one specific SNP does not benefit other SNPs of interest. We describe a simulated annealing-based algorithm that identifies an optimal selection of subjects to be phenotyped based on the weighted or unweighted average power across a group of SNPs.

Analysis Of Complex Multivariate Interactions Using Generalized Linear Latent And Mixed Modeling

✦Muhammad Yaseen, University of Nebraska, Lincoln, NE 68583, Lincoln, NE 68583, *myaseen208@gmail.com*; Kent M. Eskridge, University of Nebraska; Jose Crossa, International Maize and Wheat Improvement Center (CIMMYT)

Key Words: Multivariate Interactions, GLLAMM, AMMI, Three-way Sites Regression

Modeling complex interactions involving multiple response variables is a difficult problem in multivariate data analysis. Response variables are often correlated with each other and are influenced by factor main effects, interactions and covariates. When analyzing these types of data, the covariances and causal structure among the response variables should be taken into account. The generalized linear latent and mixed models (GLLAMMs) combine features of generalized linear mixed models and structural equation models. GLLAMM can be used to incorporate both random and fixed effects as well as model the causal structure among response variables and is more general than other methods such as AMMI and three-way sites regression which can handle only fixed effects and can't include causal structure. The use of GLLAMM to model complex multivariate interactions is illustrated with a duram wheat dataset containing seven cultivars tested over six years with four agronomic response variables and a number of covariates. The GLLAMM approach provided biologically more meaningful insight into the multivariate genotype-by-environment interactions than was possible with other methods.

FDR Doesn't Tell the Whole Story: Joint Influence of Effect Size and Covariance Structure on the Distribution of the False Discovery Proportion

✦Alan H Feiveson, NASA Johnson Space Center, Mail Code SK3, Houston, TX 77058, *alan.h.feiveson@nasa.gov*; James Fiedler, Universities Space Research Association; Robert Ploutz-Snyder, Universities Space Research Association

Key Words: false discovery, FDR, multiple comparisons, multiple testing, correlation structure

As part of a 2009 Annals of Statistics paper, Gavrilov, Benjamini, and Sarkar report results of simulations that estimated the false discovery rate (the FDR) for equally correlated test statistics using a well-known multiple-test procedure. In our study we estimate the distribution of the false discovery proportion (FDP) for the same procedure under a variety of correlation structures among multiple dependent variables in a MANOVA context. Specifically, we study the mean (FDR), skewness, kurtosis, and percentiles of the FDP distribution in the case of multiple comparisons that give rise to correlated non-central t-statistics when results at several time periods are being compared to baseline. Even if the FDR achieves its nominal value, other aspects of the distribution of the FDP depend on the interaction between signed effect sizes and correlations among variables, proportion of true nulls, and number of dependent variables. We show examples where the mean FDP (the FDR) is 10% as designed, yet there is a surprising probability of having 30% or more false discoveries. Thus, in a real experiment, the proportion of false discoveries could be quite different from the stipulated FDR.

High Dimensional Sparse Multivariate Normal Mean Testing

✦Rajarshi Mukherjee, Harvard University, 02120, *rajmrt23@gmail.com*

Key Words: Higher Criticism, Optimal detection Boundary, High Dimensional Sparse Multivariate Normal Mean Testing, Bahadur Efficiency

In High Dimensional Sparse Multivariate Normal Mean Testing problem, we have suggested a family of test statistics which, for a particular subset of the alternatives, mimics the oracle(that is the case where one knows the location of the possible signals) Likelihood Ratio Test in some "appropriate asymptotic" sense. Since the Likelihood Ratio Test, in the case one knows the exact location of the possible signals, can be considered optimal in the sense of Bahadur efficiency etc., it might seem reasonable to mimic it in "appropriate asymptotic" sense. The family of test statistics suggested is compared against the Higher Criticism Test statistic introduced by Donoho and Jin, and against the Optimal Detection Boundary as introduced by Ingster for the testing problem in question. The correlated case was also considered.

Predicting Spring Wheat Falling Number In Finland

✦Timo Hurme, University of Turku, Assistentinkatu 7, 4. krs., Turun yliopisto, 20014 Finland, *timo.hurme@utu.fi*; Pirjo Peltonen-Sainio, MTT Agrifood Research Finland; Lauri Jauhiainen, MTT Agrifood Research Finland

Key Words: falling number, prediction, web service

Timing of wheat harvest is crucial, especially in Finnish growing conditions. Early harvest results in high grain moisture content and increases the drying costs. Late harvest causes a risk of collapse in falling number - an important quality property of bread wheat - into level where the yield does not meet the standards introduced for food grain. MTT has introduced a web service for spring wheat falling number prediction. The predictions are updated daily during August and September and therefore the service offers daily fresh information on the development of falling number. The service also has an excellent geographical coverage, because the results are presented on maps covering the entire wheat production area in Finland. The falling number predictions are based on statistical models, which were developed using experimental data from years 1998-2007. The input data used to calculate the predictions from the models is the daily 10 ◊ 10 kilometer temperature, moisture and rain lattice data. In addition to the daily falling number forecast, the web service foretells the sudden crash in falling number by giving predictions for three different upcoming five day weather scenarios.

Estimating Temporal Associations In Electrocorticographic (Ecog) Time Series With First Order Pruning

✦Haley Hedlin, JHSPH Biostatistics Department, , *haleyhedlin@gmail.com*; Brian Caffo, Johns Hopkins Department of Biostatistics; Dana Boatman, Johns Hopkins Hospital, Department of Neurology

Granger causality (GC) is a statistical technique used to estimate temporal associations in multivariate time series. Many applications and extensions of GC have been proposed since its formulation by Granger in 1969. Here we control for potentially mediating or confounding

associations between time series in the context of event-related electrocorticographic (ECoG) time series. A pruning approach to remove spurious connections and simultaneously reduce the required number of estimations to fit the effective connectivity graph is proposed. Additionally, we consider the potential of adjusted GC applied to independent components as a method to explore temporal relationships between underlying source signals. Both approaches overcome limitations encountered when estimating many parameters in multivariate time-series data, an increasingly common predicament in today's brain mapping studies.

Joint Modeling Of Time-To-Event And Tumor Size

✦Weichao Bao, University of South Carolina, 247 S Marion ST, Columbia, SC 29205, *baow@email.sc.edu*; Bo Cai, University of South Carolina; Wei Shen, Eli Lilly and Company

Key Words: Joint modeling, time-to-event, longitudinal data, clinical trials, tumor size

In clinical trials, time-to-event data and longitudinal data are often collected. To model both the time-to-event and longitudinal components simultaneously, a joint modeling approach becomes increasingly important which can reduce potential biases and improve the efficiency in estimating treatment effects. In cancer clinical trials, change in tumor size is an important efficacy outcome which might serve as a surrogate for the overall survival time. In this paper, we propose a joint model where a nonlinear mixed-effect model is used to describe change in tumor size and an accelerated failure time model is used to describe overall survival time. The nonlinear mixed-effect model includes an exponential shrinkage and a linear progression. The survival and longitudinal components are linked through both fixed effects and random effects with appropriate adjustments. A simulation study is presented to evaluate the performance of the proposed approach compared with other competing models.

Study Design And Analysis By Means Of Free-Response Roc (Froc) Data; An Evaluation Of Clinical Radiological Data

✦Takayuki Abe, Keio University School of Medicine, 35 Shinanomachi Shinjuku-ku, Tokyo, International 160-8582 Japan, *tabe@z5.keio.jp*; Yuji Sato, Keio University School of Medicine; Yoshitake Yamada, Keio University School of Medicine; Kenji Ogawa, Nippon Koukan Hospital; Sachio Kuribayashi, Keio University School of Medicine; Manabu Iwasaki, Seikei University

Key Words: ROC, free-response ROC (FROC), jackknife alternative FROC (JAFROC)

The receiver operating characteristic (ROC) method is commonly used in observer performance studies for medical diagnostic devices where every subject is given a score as a diagnostic result. On the other hand, localizing and marking multiple abnormalities on an image is sometimes clinically warranted and, as a result, some images can be given more than one score. Such data is called free-response data, and corresponding FROC curve is defined. One of the frequently applied methods for analyzing such data is the jackknife alternative FROC (JAFROC) method where an AUC-like summary index is used. In this study, an evaluation was made on the effect of the selection of mixed-effects models in the JAFROC when comparing two correlated FROC curves adjusted for multiple observers. It was shown that parsimonious models tend to have higher power in the comparison. Efficiency of some study designs (including balanced incomplete block (BIB) design) was also evaluated. A real example is presented, and used in the evaluation, from a clinical research comparing two diagnostic imaging devices for the detection of pulmonary nodules.

An Empirical Maximum Likelihood Roc Model With Monotonic Likelihood Ratio

✦lucas simplice tcheuko, University of Maryland college park, 3554 childress terrace, Burtonsville, MD 20866, *lucast@umd.edu*

Key Words: empirical likelihood, convex, PAVA, constrained estimates

Although an ideal ROC curves is expected to be convex, published fits often displays ``hooks".This article presents a method of computing an empirical likelihood estimator of the ROC assuming convexity.For a given set of binary data, we first derive its empirical ROC without convexity assumption, and then derive the analytical convex maximum likelihood estimator of the ROC. We use use PAVA -Pool Adjacent Violator Algorithm- to compute the real values of the convex ROC estimator and then overlay the convex plot on the empirical plot.

Group Testing Regression Models For Multiple Traits

✦Boan Zhang, University of Nebraska-Lincoln, Lincoln, NE 68583, *boan.zhang@huskers.unl.edu*; Christopher Bilder, University of Nebraska-Lincoln; Joshua M Tebbs, University of South Carolina

Key Words: correlated binary data, expectation-solution algorithm, generalized estimating equations, latent response, pooled testing, unobserved response

Group testing, where groups of individual specimens are composited to test for a binary trait (e.g., infectious disease), is a procedure commonly used to reduce the costs of screening a large number of individuals. Group testing data is unique in that only group responses may be observed, but inferences are necessary at the individual level. A further challenge arises when specimens are screened for multiple traits leading to unobserved correlated binary responses for the individuals. In our poster, we propose the first regression models for these types of responses. Through the use of generalized estimating equations and the expectation-solution algorithm, we develop methodology to fit models when only the group responses are available. An important consequence of our methodology is that the proposed regression models are easily adapted to a longitudinal testing situation where individual subjects appear in the same groups over time. Simulation studies are performed to evaluate and compare small sample performance of our methods. Finally, the proposed modeling procedures are applied to chlamydia and gonorrhea screening data collected as part of the Infertility Prevention Project.

Adjusted Quadratic Inference Functions (Qif) And Adjusted Adaptive Estimating Equations For Fitting Marginal Regression Models For Longitudinal Data With Time-Dependent Covariate

✦Yi Zhou, Tulane University, 27560 USA, *joeyzhou2009@gmail.com*; John Lefante, Tulane University; Rice Janet, Tulane University; Shande Chen, University of Northy Texas Health Science Center

Key Words: marginal regression, estimation equation, time dependent

Statistical methods in analyzing longitudinal data with time-dependent covariates are limited. Recently some methods such as Qudratic Inference Function (QIF)and Adaptive Estimation Equation (AEE) were introduced for analyzing longitudinal data (Qu(2000,2003))and showed appealing features. However,for longitudinal data with time dependent covariates, both methods have significant limitations in terms of bias and efficiency. Thus, we propose adjusted QIF and AEE for fitting marginal regression models for longitudinal data with time-dependent covariates by restricting the moment conditions to those valid for time-dependent covariates. Our simulation studies show that both bias and efficiency of estimators using adjusted QIF and AEE are significantly improved compared to estimators using traditional QIF and AEE. In addition, we apply adjusted QIF and AEE approaches to anthropometric screening study and compared the bias and efficiency of generalized method of moments (GMM) which was introduced by Lai and Small (2007).

Goodness-Of-Fit Tests For Logistic Regression

✦Sutan Wu, Florida State University, Tallahassee, FL 32306, *titiwu@gmail.com*; Dan McGee, Florida State University

Key Words: goodness-of-fit test, logistic regression, generalized linear model

The generalized linear model and particularly the logistic model are widely used in public health, medicine, and epidemiology. Goodness-of-fit tests for these models are popularly used to describe how well a proposed model fits a set of observations. These different goodness-of-fit tests all have individual advantages and disadvantages. In this poster, we mainly consider the performance of the ``Hosmer-Lemeshow" test, the Pearson's chi-square test, the unweighted sum of squares test and the cumulative residual test. We examined their performance in a series of empirical studies as well as simulation scenarios. We conclude that the cumulative sums of residuals test gives better overall performance than the other three. We also conclude that the commonly suggested practice of assuming that a p-value less than 0.15 is an indication of lack of fit at the initial steps of model diagnostics should be adopted. Additionally, D'Agostino et al. presented the relationship of the stacked logistic regression and the Cox regression model in the Framingham Heart Study. So in our future study, we will examine the possibility and feasibility of the adaption these goodness-of-fit tests to the Cox pr

Reconstructing Dna Copy Number By Joint Segmentation Of Multiple Sequences

✦Zhongyang Zhang, UCLA, 76 Barnes Ct Apt 116, Stanford, CA 94305, *zhangzy@ucla.edu*; Kenneth Lange, University of California, Los Angeles; Chiara Sabatti, Stanford University

Key Words: copy number variant, group fused lasso, MM algorithm, joint segmentation

DNA copy number variants (CNV) are gains and losses of DNA segments as compared to normal counterparts. Recent results have revealed that CNV comprises a ubiquitous and important class of structure variants among the full spectrum of genetic variants, and underscored its role in explaining complex disease and other phenotypic traits. A number of high-throughput technologies provide signal that can be used to reconstruct DNA copy number and numerous methods for their analysis have been developed. We tackle the problem of joint analysis of multiple signals, which are expected to accumulate evidence on consensus copy number states. This important development is motivated by a variety of practical situations. We proposed a join segmentation approach based on group fused-lasso. A majorization-minimization strategy is employed to attack this non-trivial optimization problem, complicated further by the heterogeneity of data structure. The computation cost is linear in the number of probes times the number of samples. We illustrate the results with simulation and the analysis of pedigrees collected to study the genetic underpinning of bipolar disorder.

Biomarker Discovery In Lipidomics - A Systematic Comparison Of Methods

✦Maiju Elisa Kujala, University of Turku, Assistentinkatu 7, Department of Social Research, Turun yliopisto, International 20014 Finland, *mekuja@utu.fi*

Key Words: Lipidomics, variable selection, biomarker detection

Lipidomic analyses, among other ''omics'' data sets, are becoming a popular part of pharmaceutical drug development. The goal of the present project is to develop statistical methodology for high dimensional and dependent data with real-life applications in lipidomics data sets keeping in mind that methods used should be acceptable both from scientific and regulatory points of view. The current research is a systematic comparison of methods used in biomarker discovery such as stepwise logistic regression, lasso logistic regression, and (regularized) discriminant analysis, among others. The ability of recover structures in these high dimensional data sets is studied by extensive simulations as well as illustrative examples. In addition, special issues such as missing values, sources of variation and respective normalizations and strong correlations between lipids are addresses.

A Bootstrap Approach For Testing Marginal Independence Between Two Categorical Variables When Subjects Have Repeated Responses

✦Rhonda J Rosychuk, University of Alberta, , *rhonda.rosychuk@ualberta.ca*; Christina Alloway, ; Amanda S Newton, University of Alberta

Key Words: non-parametric bootstrap, contingency table, marginal independence, correlated data, categorical data

Two-way contingency tables are used to classify subjects by two categorical variables. To assess independence in these tables, the Pearson chi-square test or Fisher's exact test are typically used. These tests assume that each subject contributes at most one count to only one table cell (e.g., sex versus blood type). In other situations, each subject may

have more than one count contributing to the table. One may wish to test independence, adjusting for the within-subject correlation. We provide a simple non-parametric bootstrap approach and assess its performance through simulation studies. The method is illustrated on subjects with multiple mental health presentations to Emergency Departments.

360 Contributed Oral Poster Presentations: Business and Economic Statistics Section

Business and Economic Statistics Section

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

The Effect of the Business Cycle on the Performance of Socially Responsible Equity Mutual Funds

✦Andrea Roofe, Florida International University, c/o 3581 SW 117th Avenue, Apt 206, Miami, FL 33175, *andrea.roofe@fiu.edu*

Key Words: switching regression, business cycle, finance, mutual fund, ecomometrics, recession

The paper applies the switching regression originally proposed by Goldfeld & Quandt (1973) to examine the behavior of a sample of 10 socially responsible equity mutual funds during the expansion and contraction phases of US business cycles between July 1991 and June 2009, the end of what is often called The Great Recession, using monthly data. The information on business cycles was tracked using the dividend yield. Fund returns were less volatile during an expansion than during a period of contraction, as indicated by the standard deviation of returns. The Carhart 4-factor model explained 70% of the variability in fund excess returns. During an economic expansion, fund excess returns were explained by the differential between high and low ratios of book to market value (HML) and market excess returns. During an economic contraction, fund excess returns were explained by the differential between small and big firms (SMB), momentum, and market excess returns.

Segmenting The Time Series Of Quarterly Gdp Using A Hidden Markov Model

✦Yu Chen, Liautaud Graduate School of Business, University of Illinois at Chicago, IDS DEPT (MC 294), 601 S MORGAN ST, CHICAGO, IL 60607-7124, *ychen31@uic.edu*; Stanley L Sclove, Information & Decision Sciences Dept., Univ. of Illinois at Chicago

Key Words: time series, segmentation, GDP, hidden Markov model

The time series of quarterly growth rates of US GDP, from 1947 through the third quarter of 2010, was segmented by hidden Markov models (HMMs). HMMs with several states were fit and compared with a single distribution for the growth rate. State-conditional Normal distributions with different means and variances were fit for different numbers of states. The extent to which states correspond to recession, recovery, expansion, and contraction was assessed. The HMMs were scored by BIC. Some comparison was made to ARIMA models involving regular and quarterly differencing and regular and quarterly autoregression of log GDP. Components of GDP were also fit with HMMs, with a view toward determining which components are leading or lagging indicators of the state of overall GDP.

Sieve Bootstrap Prediction Intervals For Multivariate Arma Processes With Non-Gaussian Innovations

✦Purna Mukhopadhyay, Univeristy of Kansas Medical Center, 5620 W 133rd Terrace Apt 611, Overland Park, KS 66209, *pmukhopadhyay@kumc.edu*; V A Samaranayake, Missouri university of Science and Technology

Key Words: Sieve Bootstrap, Multivariate, ARMA processes, non-Gaussian, nonparametric

Existing nonparametric bootstrap methods for obtaining prediction intervals for vector autoregressive moving average (ARMA) processes require apriori knowledge of the autoregressive and moving average orders, p, q respectively. The sieve bootstrap method developed for stationary and invertible univariate processes overcomes this limitation. We implement a modified version of the sieve bootstrap method to multivariate ARMA processes and show, through a Monte Carlo study, that the procedure produces prediction intervals that achieve nominal or near nominal coverage probabilities. The robustness of this method under non-normality is tested using different error distributions and Monte Carlo results show that the coverage remains at nominal or near nominal levels under several non-normal distributions.

Using Repeated Measures To Estimate Elasticities Of Demand For Retail Electricity In The State Of Michigan

✦Michael Taylor, ITC Holdings Corporation, 27175 Energy Way, Novi, MI 48377, *mtaylor@itctransco.com*

Key Words: Energy Price Elasticity, Utility Economics, Weather Sensitivity of Energy Demand, Electrical Demand, Price Elasticity of Demand

In forecasting retail electrical sales, the estimation of elasticity of demand, particularly with respect to price, has long been a source of controversy among utilities and regulators. Using a log-log, mixed effect model, we derive significant estimates for elasticity of demand with respect to price, weather and economics. These are treated as fixed effects, where a 'subject' (electric utility) variable is entered as a random effect, controlling for the homogeneity within subjects (i.e. annual consumption per customer for a given utility). The model was fit on a sample of Michigan electrical utilities over the period 2002-2007. Our price elasticity estimates showed that the larger the consumption profile of a retail customer class, the more sensitive their consumption was to price movements. This modeling approach could be applied to a variety of contexts where a seller of a product sells a similar product to similar sets of customers repeatedly, and wishes to analyze the sensitivity of their consumption with respect to a set of covariates. Examples would include energy, foodstuffs, sports/entertainment tickets, charitable contributions and rental rates of real property.

Efficient Quantile Regression For Linear Heterogeneous Models

✦Yoonsuh Jung, MD Anderson Cancer Center, 3720 W. Alabama st, apt 3203, Houston, TX 77027, *yjung1@mdanderson.org*; Yoonkyung Lee, Ohio State University; Steven N MacEachern, Ohio State University

Key Words: check loss function, heteroscedasticity, quantile regression

Quantile regression provides estimates of a range of conditional quantiles. This stands in contrast to traditional regression techniques, which focus on a single conditional mean function. Lee et al. (2009) proposed efficient quantile regression by rounding the sharp corner of the loss. The main modification generally involves an asymmetric L2 adjustment of the loss function around zero. The adjustment leads to superior finite sample performance by exploiting the bias-variance tradeoff. We extend the idea of L2 adjusted quantile regression to two linear heterogeneous models. The first model involves a set of weights in quantile regression (as in one description of weighted least squares). The second incorporates location-scale model, which introduces different weights and is preferable due to an invariance argument. We discuss several choices of reasonable weights that can be practically useful. Finally, the L2 adjustment is constructed to diminish as sample size grows. Conditions to retain consistency properties are provided.

The Blup'S Asymptotic Distribution: A Didactical Note.

✦Luis Frank, University of Buenos Aires, Av. San Martin 4453, Buenos Aires, International C1417DSE Argentina, *lfrank@agro.uba.ar*

Key Words: BLUP, Mixed Models, Linear Models, Random Effects

The paper shows in a simple way that the variance of a mixed-model's BLUP (obtained by the least squares criterion) converges asymptotically to the variance of a fixed-parameters estimator. As a consequence, the asymptotic distribution of the random-effects predictor is also similar to the distribution of the fixed-effects estimator, a result that seems to have been overlooked in the literature.

Limit Theorems For Strictly Stationary Random Fields Satisfying A Strong Mixing Condition

✦Cristina Tone, University of Louisville, 328 Natural Sciences Building, University of Louisville, Louisville, KY 40292 U.S., *crtone@indiana.edu*

Key Words: limit theorems, empirical processes, strong mixing condition, Gaussian process, random fields, Hilbert space

We introduce some limit theorems for strictly stationary random fields satisfying an interlaced mixing condition. We proceed by first presenting a common technique in proving limit theorems for real- valued dependent random sequences, followed by some extensions to the case of Hilbert-space valued random fields, and empirical processes endowed with real values from a strictly stationary random field satisfying a certain mixing condition.

Seasonal Stochastic Volatility Models

✦Julieta Frank, University of Manitoba, 66-353 Dafoe Rd., Winnipeg, MB R3T 2N2 Canada, *julietafrank@yahoo.com*; Melody Ghahramani, University of Winnipeg; Aerambamoorthy Thavaneswaran, University of Manitoba

Key Words: seasonality, stochastic volatility, kurtosis, forecast error variance

New information entering into the market causes changes in price volatility. Often these changes occur at regular time intervals, for example, at market opening and closing times, at certain times during the day, at specific days of the week, or around weekends and vacation periods. Even though much research has been done on volatility models, more general specifications accounting for seasonal volatility have been little explored. In this paper, we focus on the random coefficient autoregressive models with seasonal stochastic volatility errors. We derive the moments of the error distribution and the closed-form expression for the variance of the l-steps ahead forecast error. The results are a generalization of the non-seasonal version of the model. The expressions derived here can be used to obtain the moment estimates of the model parameters and hence more precise estimates and better forecasts of market behavior, and to assist investors, decision makers, and other market participants in developing trading strategies.

Measures Of The Economic Value Of Bankruptcy Probabilities

David Johnstone, University of Sydney; Stewart Jones, University of Sydney; Maurice Peat, University of Sydney; ✦Victor Richmond Jose, Georgetown University, Washington, DC 20057, *vrj2@georgetown.edu*

Key Words: Scoring Rules, Forecast Evaluation, Bankruptcy Probability, Default Risk

Financial institutions and regulatory agencies direct much effort and expertise towards estimating bankruptcy probabilities. By comparison, the techniques used to evaluate probability estimates have attracted little attention and still remain mostly ad hoc. We introduce a family of economic probability score functions designed to capture the utility obtained by a user, with a specified utility function, who uses the estimated probabilities to make hypothetical bets against a rival forecaster or model. The conceptual appeal of these statistical score functions is that probability forecasts are evaluated neither in abstract, nor in isolation, but instead by whether they would have "made money" for a given user, with specified risk aversion, against comparable forecasts or market betting prices.

M-Stationary (Multiplicative Stationary) Processes- A Revisit

✦Md Jobayer Hossain, Nemours, 1701 Rockland Road, Wilmington, DE 19803 USA, *mhossain@udel.edu*; Wayne A Woodward, Southern Methodist University; Henry L Gray, Southern Methodist University

Key Words: M-stationary, elongating, compacting, time varying, periodic

M-stationary processes, introduced by Gray and Zhang, were a breakthrough in the analysis of non-stationary time series whose pseudo periodic behavior lengthens approximately linearly with time. They used a time deformation technique for the analysis of continuous M-stationary processes with a time index (t) restricted to (0,Inf). Gray, Vijverberg and Woodward extended these results to discrete processes. Later all these results were generalized by the introduction of G(Lambda) processes. Euler processes and Harmonic in Log Space processes are examples of M-stationary processes. These methods outperformed the conventional methods. In the current research, we extend the definition of M-stationary processes on t with the ranges (-Inf,0) and (0,Inf) for analyzing non-stationary processes with periodic behavior that compresses or lengthens with time. We refer to M-stationary processes as Left M-stationary and Right M-stationary when t is restricted to (-Inf,0) and (0,Inf) respectively. The new M-stationary processes produced superior results compared to conventional approaches for analyzing time series with an approximate linear compacting or elongating pseudo periodic behavior.

The Properties And Effectiveness Of Filter Trading Rule

✦Ling Xin, The University of Hong Kong, Room 518, Meng Wah Complex, The University of Hong Kong,, Hong Kong, Hong Kong, China, *xl.elaine@gmail.com*

Key Words: filter rule, Markov switching model

Filter trading rule generates a sequence of buy/sell signals according to the following principle. If the asset price moves up at least 100d% from a low, the signal sequence will start with a buy. We then buy and hold the asset until the closing price moves down at least 100d% from a subsequent high, at which time a sell signal is generated and we simultaneously sell and go short. Here d is called the filter size of the filter trading rule. In this paper, we discuss the duration and profitability of filter trading rule and explore the problem of finding a suitable filter size to maximize the trading profit. We discussed existing contributions which are concentrated on random walk framework and then extend to Markov switching model. We constructed an absorbing Markov chain as well as a stationary Markov chain to separately study the durations and profits of filter trading under Markov switching markets. Numerical example under a market with mixture Bernoulli distributions is illustrated. The continuous time Markov switching model leads integral equations for the quantities in our problem.

Binary Prediction To Minimize Total Risk

Kentaro Akashi, Institute of Statistical Mathematics; ✦Yoshinori Kawasaki, Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo, International 1908562 Japan, *kawasaki@ism.ac.jp*

Key Words: binary prediction, risk minimization, optimal cutoff-point, minimum prospective interval, asymptotic theory

Risk involved with financial contracts often can be viewed as uncertainty of binary outcomes. This paper treat risk minimization as the problem of profit maximization, and gives an optimal solution of the cut-off point for binary prediction. This optimality or profit maximization will be asymptotically attained in the sense of convergence in probability. In practice, we have to replace the true parameters inside an indicator function by their estimates. Because indicator functions are discontinuous, apparently it looks non-standard argument. We show, in spite of this, that we can construct an interval for maximized profit, and even minimize it based on asymptotic theories where the MLE is simply plugged in. Simulation results suggest that the finite sample properties of our asymptotic theories are satisfactory. In an empirical analysis using personal loan data of a south German bank, we show the total profit realized by our optimal prediction exceeds the actually observed profit regardless of the settings of loan interest.

Portfolio Management Strategies

Les Yen, University of Phoenix/District of Columbia; ✦Jonathan Hale, University of Phoenix/District of Columbia, , *jjhale7@yahoo.com*

Key Words: Portfolio management, Diversification, Rebalancing strategies, Economic indicators, Business cycles, Correlation

Most people accept the investment concept to diversify in portfolio management and construction. This paper examines the market environment, economic indicators and conditional correlations that influence the outcome of strategies, decisions and investment performance. We show a portfolio management strategy that exhibits superior returns.

Comparison Of Forecasting Approaches Using Proc Reg And Proc Arima Vs. Sas Time Series Forecasting System

✦Martin Selzer, Chatham Decision Sciences, 72 Hedges Avenue, Chatham, NJ 07928, *mselzer@att.net*

Key Words: SAS Macros, Times Series, Regression, Arima, Forecasting

This paper describes a SAS macro based program using Proc Reg and Proc Arima developed for forecasting. The program is compared to a separate forecasting process using the SAS Time Series Forecasting System (SASTSFS). Forecasts are prepared with the macro based program and compared to those produced by other programs and SASTSFS. Results show that the new program gives results quite near those generated by SASTSFS. While no effort was made to compare the accuracy of the two approaches against actual results, the mathematical formulation of the statistical models used in the macro based program suggest the program would provide more accurate goals. The macro based approach also has fewer manual interventions and produces forecasts more quickly and less prone to error.

Prediction Of Stock Price Based On Forward And Backward Stochastic Differential Equation

✦Yun Zheng, Iowa State University, 116 20th St, Ames, IA 50010 United States, *yunzheng@iastate.edu*

Key Words: Stochastic Functinal Differential Equation, Black Scholes formula, Option Pricing, Anticipating Calculus, Insider Trading

An insider is an agent who has potential access to non-public information. In this paper, we compute the logarithmic utility of an insider when the financial market is modeled by a forward and backward stochatic differential equation. Further, an optimal trading strategy for insider is given.

361 Contributed Oral Poster Presentations: International Chinese Statistical Association

International Chinese Statistical Association

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Finding Transcription Factors Of Co-Regulated Genes

✦Dongseok Choi, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, OR 97239 USA, *choid@ohsu.edu*; Kathryn Carr, Oregon Health & Science University ; Lauren Hayashi, Oregon Health & Science University ; Ted Acott, Oregon Health & Science University

Key Words: transcription factor, microarrays, binding sites, glaucoma

Transcription factors do a crucial role in regulating gene expression patterns. Transcription factor binding sites (TFBS) can be found based on database searches of known transcription factor binding sites or by using statistical models. In this presentation, we will compare results from both approaches and combine them for microarray array experiments of glaucoma treatments.

Profile Monitoring Of Batch Processes With Pi Controllers In Semiconductor Manufacturing

✦Shui-Pin Lee, Ching Yun University, No. 229, Jianxing Road, Zhongli City,, Taoyuan County, Zhongli, 320 Taiwan, *shuipin@cyu.edu.tw*

Key Words: profile monitoring, batch process, proportional-integral controller, health index, state space, average run length

A modern semiconductor manufacturing line contains hundreds of sequential batch processing stages. Each of these operation stages consists of many steps carried out by expensive tools with several PI controllers to adjust the recipes, which are monitored by numerous sensors capable of sampling at intervals of seconds. The heterogeneous variations at different profile points are mainly due to on-off recipe actions of the PI controllers at specific points. In addition, the analysis of these profiles is further complicated by long-term trends due to tool aging and short-term effects specific to the first wafer in a lot-cycle. Statistical process control methods that fail to take these effects into consideration will lead to frequent false alarms. A systematic method is proposed to address these challenges. At first, each of these PI controllers is described by a state space model, its output matrix is used to capture and remove intrinsic variations due to long-term aging trends and the short-term first-wafer effects. The residuals are used to formulate a health index, and this index can be used to monitor the health of the equipment and detect faulty wafers efficiently.

362 Contributed Oral Poster Presentation: Section on Government Statistics

Section on Government Statistics

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Obtaining Missing Race For Chronic Kidney Disease (Ckd) Patients Using Multiple Department Of Veterans Affairs (Va) Administrative Databases

✦Hong Jen Yu, Department of Veterans Affairs, Houston HSR&D Center of Excellence, 2002 Holcombe Blvd. (152), Houston, TX 77030, *HongJen.Yu@va.gov*; Nancy J. Petersen, Department of Veterans Affairs; G. John Chen, Department of Veterans Affairs

Key Words: Administrative databases, race, data quality

Background: A large proportion of Veterans who receive care in VA have missing race information if race is determined using a single clinical data source. VA administrative databases include inpatient, outpatient, pharmacy, lab and cost data. Race is included in some databases but it may be missing for many patients. For a study examining racial differences in services in CKD patients derived from a lab dataset that did not contain race, we obtained race using inpatient and outpatient files. Methods: We determined the percent of patients with missing values for race in the inpatient and outpatient datasets from fiscal years (FYs) 1997-2010. We used an algorithm to look at all years simultaneously to extract any values of race that might be in the files. Results: The percent missing race ranged from 42.8% in FY97 to 20.8% in FY10 for outpatient files and 2.9% to 56.0% for inpatient files. After combining files across all years, we obtained race on 84.5% of the 1.2 million CKD patients. Conclusions: Examining multiple years simultaneously allowed us to obtain race for a high percent of our study patients that would have been missing if we used a single clinical database.

Method Comparison For Assessing Trends Over Time Of Age Of First Cigarette Use Among High School Students In The U.S. - Youth Risk Behavior Survey, 1991-2009

✦Emily O'Malley Olsen, CDC/NCCDPHP/DASH, 4770 Buford Highway, Mailstop K-33, Atlanta, GA 30341, *dgx1@cdc.gov*; Sherry Everett Jones, CDC/NCCDPHP/DASH

Key Words: trend analysis, surveillance, Youth Risk Behavior Survey, adolescent health, smoking

Secular trend analysis of Youth Risk Behavior Survey (YRBS) data identifies changes in prevalence of youth cigarette use over time. A single YRBS question assessing age of first whole cigarette use captures both the prevalence of smoking and the age of first cigarette use, both of which have important public health implications. To best understand trends in cigarette smoking among high school students nationwide, this question was analyzed using multiple modeling approaches: time-until-event outcome "years from birth until first cigarette" modeled with Cox PH regression, a linear outcome "age of first cigarette" modeled with linear regression, and dichotomous outcomes "smoked

before age 13" and "ever smoked" modeled with logistic, Poisson, and negative binomial regressions. Different modeling approaches might result in different conclusions. Conversely, they might result in similar conclusions but might be more computationally intensive and difficult to interpret. Comparisons between each of the different modeling approaches and their resulting public health implications will be presented.

Controlling For Multiple Testing In An Investigation Of The Association Between Occupation And Mortality From Diabetes

✦Jia Li, NIOSH, 4676 Columbia Parkway, MS R-17, Cincinnati, OH 45226, *qzl0@cdc.gov*; Cynthia F Robinson, NIOSH; James T Walker, NIOSH

Key Words: multiple testing, familywise error rate, false discovery rate

Multiple testing problems arise frequently in statistical data analysis. Choice of the appropriate procedure to correct for these problems is important in order to avoid erroneous inferences. The National Occupational Mortality Surveillance System (NOMS) contains death certificate data from up to 28 states for the years 1984-1998 with coded information about usual occupation. In this study, we used data from the NOMS system to examine the association between occupation, gender, and death from diabetes. The problem of multiple testing is relevant, because occupation was coded in detailed categories and inferences were made simultaneously for the set of categories. In the present work, we illustrated the use of the false discovery rate (FDR) approach to the multiplicity problem with three examples. We addressed the differences in interpretation between familywise error rate (FWE) and FDR, and illustrated use of FDR control in the case of both independent tests and tests under dependency. We concluded that the FDR approach is appropriate for exploratory data analysis with a large number of tests to maintain a balance between type I and type II errors.

The Weighted Brennan-Prediger Coefficient For Multiple Raters In Low Variance Data With Missing Values

✦Mario Jose Nunez, IDA Science and Technology Policy Institute, 1899 Pennsylvania Ave., NW, Suite 520, Washington, DC 20006, *mnunez@ida.org*; Jamie Doyle, IDA Science and Technology Policy Institute; Amy Marshall Richards, IDA Science and Technology Policy Institute; Mary Beth Hughes, IDA Science and Technology Policy Institute

Key Words: Inter-rater, agreement, kappa, Brennan-Prediger, reliability, interval

In 2007, the IDA Science and Technology Policy Institute (STPI) was commissioned to evaluate the National Institutes of Health Director's New Innovator Award. As part of the evaluation, STPI assessed the application review process, including the degree of inter-rater reliability, in order to determine whether the program was implemented according to its goals. During the review process, applications were judged by three external reviewers on four criteria. After testing several widely used inter-rater reliability measures, STPI found the measure of reviewer agreement to be fairly low, which was counter-intuitive given the apparent strong agreement in the data. This paper investigates the history of various inter-rater reliability measures and their specific applications to different data structures. As part of this study, a sensitivity analysis was conducted to compare common agreement coefficients. Based on our findings from evaluating the award, we suggest a modified version of the Brennan-Prediger (BP) Coefficient (1981) which can manage missing data. We argue this to be a dependable coefficient when data are missing at random, have a small range, and are on an interval scale.

How Large Is The Finance Company Universe? A Unique Challenge And A Solution

✦Lisa Chen, Federal Reserve Board, 20th & C St. NW, Washington, DC 20551, *lisa.x.chen@frb.gov*

Key Words: Cluster sampling, Estimation, Nonresponse, Business survey

The Census of Finance Companies was conducted in 2010 by the Federal Reserve Board to assess companies that supply credit or lease financing to households or businesses. Nonresponse follow-up activities were also conducted as part of an effort to estimate the universe. One unique challenge was the inadequate sampling frame which turned out to contain records including branches or offices of companies instead of the likely financial companies deemed to be the highest holder in their hierarchical corporate structure. This paper presents a clustering approach based on the similarity in company names for the design and implementation of the nonresponse follow-up sample. Furthermore, this paper also discusses the impact on and the solution for producing universe estimates as well as the effectiveness of the sample design.

363 Contributed Oral Poster Presentations: Section on Statistics and the Environment

Section on Statistics and the Environment

Tuesday, August 2, 10:30 a.m.–12:20 p.m.

Air Quality Assessment Survey in Lagos Metropolitan City and Its Implication on Nigerian MDGs' Environmental Statistical Indicators

✦Ismaila Adeleke, University of Lagos, Department of Actuarial Science and Insurance, Faculty of Business Administration, Lagos, 1014 Nigeria, *adeleke22000@yahoo.ca*; Hamadu Dallah, University of Lagos; Olukemi Odukoya, University of Lagos; Shakirudeen Odunuga, University of Lagos; Aderonke Oyeyiola, University of Lagos; Ismaila Adeleke, University of Lagos; Ismaila Adeleke, University of Lagos; Raymond Okafor, University of Lagos; Shakirudeen Odunuga, University of Lagos

Key Words: Air Quality Assessment, Environmental Data, Survey Methodology, Pollutants, GIS, MDGs

Protecting public health and the environment has stimulated a lot of attention in recent times. Environment is one of the common pillars of the Millennium Development Goals (MDGs) due to its global impact. One of the most complex and important global tasks for the future of the planet is to protect and sustain the natural environment. However, meeting this target looks like a mirage for developing countries like Ni-

geria mainly due to lack of environmental data and information. The Federal Ministry of Environment (FMEnv) set limits for ambient concentrations of six "criteria" pollutants (Carbon Monoxide (CO), Lead, Nitrogen Dioxide (NO2), Sulfur Dioxide (SO2), Particulate Matter, and Noise) considered harmful to the environment and public health. Recognizing that knowledge about health and environmental impacts assessment of air quality pollutants should necessitate regular/periodic air quality surveys. However, the geometrical increase in the use of power generators in the target environment in recent years called for urgent need for air quality research to ascertain the current levels of the six pollutants in order to establish its potential health and economic impacts.

Microbial Diversity Of Anaerobic Granules In Wastewater Treatments

✦Bo Hu, University of Minnesota Twin Cities, Saint Paul, MN 55108, *bhu@umn.edu*; Wei Wei, University of Minnesota Rochester

Key Words: Richness, Abundance, Clone Library, Wastewater

Studies have shown that the microorganisms in anaerobic granules from wastewater can be treated and utilized to produce biohydrogen, an important bioenergy carrier. Treatments of anaerobic granular with chloroform were reported as efficient ways in transforming wastewater to a hydrogen-producing system. However, there is no statistical analysis about whether or not the microbial community in anaerobic granules, including richness, abundance, evenness and species, changes after the chloroform treatment and, furthermore, if the changes contribute to hydrogen production. Prokaryotic genomic DNA from anaerobic granule samples, in wastewater, before and after the chloroform treatment were extracted and 16s rRNA gene clone libraries were constructed for both bacteria and archaea separately from the two samples. Statistical methods including Chao's estimation, rarefaction curve, Shannon's index and bootstrap estimation were applied to estimate the richness, abundance and evenness of bacteria and archaea clone libraries of these two samples. We found significant decreases of richness and evenness after the chloroform treatment for both bacteria and archaea clone libraries.

Statistical Modeling Of Environmental Conditions In Paint Creek

✦Robert Kushler, Oakland University, 24228 Edgemont Dr, Southfield, MI 48033-6425, *kushler@oakland.edu*; Scott Tiegs, Oakland University; Adam Avery, Oakland University

Key Words: stream ecology, graphics, nonlinear models

Environmental conditions in Paint Creek, a tributary of the Clinton River in southeast Michigan, have been monitored over the past two years. Large amounts of data on temperature, water depth, and other conditions were captured. Some non-standard statistical techniques that were used to analyze the data and assess the accuracy of a mathematical model for stream flow will be illustrated.

Statistical Assessment Methods For Ambient Water Quality Criteria For Dissolved Oxygen

✦Koji Kanefuji, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo, 190-8562 Japan, *kanefuji@ism.ac.jp*; Kosei Iwase, Yokohama College of Pharmacy; Mitsumasa Okada, The Open University of Japan

Key Words: water quality criteria, dissolved oxygen

In the Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll a for the Chesapeake Bay and Its Tidal Tributaries (U.S. EPA 2007), the cumulative frequency diagram (CFD) approach was introduced. The CFD-based water quality criteria assessment methodology offers the potential for significant benefit in assessing the target bay water quality criteria attainment. We checked up some statistical properties of this method. In the result, redefinition of the CFD-based water quality criteria assessment methodology is implemented in an effective and efficient manner. We present some relevant statistical methods of the CFD-based water quality criteria assessment. Some analytic results for the Bay data in Japan are represented in this study.

On The Asymptotics Of Maximum Likelihood Estimation For Spatial-Temporal Linear Models On A Lattice

✦Xiang Zhang, University of Kentucky, 40508, *xiang.zhang2@uky.edu*; Yanbing Zheng, University of Kentucky

Key Words: Autoregressive models, increasing domain asymptotics, linear regression, spatial-temporal process

spatial-temporal linear model and the corresponding likelihood-based statistical inference are important tools for the analysis of spatial-temporal lattice data and have been applied in a wide range of disciplines. However, understanding of the asymptotic properties of maximum likelihood estimates under general asymptotic framework is limited. Here we focus on spatial-temporal simultaneous autoregressive models. Under increasing domain asymptotic framework, we propose mild regularity conditions on the spatial-temporal weight matrices and derive the asymptotic properties (consistency and asymptotic normality) of maximum likelihood estimates. A simulation study is conducted to examine the finite-sample properties of the maximum likelihood estimates.

Em Algorithm For Matching Multiband Satellite Data And Land-Cover Classification

✦Jason Stover, Georgia College, Mathematics Department, CBX 017, Milledgeville, GA 31061, *jason.stover@gcsu.edu*; Matthew Ulm, Georgia College; Timothy Andrzejewski, Georgia College; Matthew Yonz, Georgia College

Key Words: iterated conditional modes, remote sensing, markov random field

Matching reflectivity of infrared Earth images of different resolutions is possible with the EM algorithm. These matched data can then be used to estimate land cover via iterated conditional modes. Data from NASA's ASTER mission consist of reflectivity o f Earth's surface measured in nine different wavelengths. Each wavelength has a resolution of either 15m or 90m, which makes simultaneous use of all bands difficult given the consequent mismatch of pixels. Using the correlation among bands and training data taken from the ground, the EM algorithm was used to impute the reflectivity of the larger pixels, thereby improving the estimated land-cover.

An Anchor Placement Approach In The Method Of Anchored Distributions

✦Yarong Yang, University of California Berkeley, Berkeley, CA 94720 USA, *yarongyang78@berkeley.edu*; Matt Over, University of California Berkeley; Haruko Murakami, University of California Berkeley; Yoram Rubin, University of California Berkeley

Key Words: inversion, Singular Value Decomposition, likelihood, Bayesian, sensitivity

The method of anchored distributions (MAD) is a general Bayesian inversion technique aimed at estimating the parameters in distributed-parameter fields. Anchors, the central element of MAD, are statistical distributions of the target parameters at specific locations, which are used to localize large-scale, indirect data. They are intended to capture the information contained in multi-type, multi-scale data that is relevant for the inversion and express it in terms of the dependent variables. It is important to work with a small number of anchors in order to reduce the dimensionality of the likelihood function, and to achieve that, anchors must be placed strategically. In this study we employ Singular Value Decomposition (SVD) of the sensitivity matrix, elements of which express the sensitivity of each data location to each potential anchor location, to identify such strategic locations. The locations selected by our proposed method are tested in synthetic studies. Comparison studies between inversion based on anchors placed at strategic locations vs. anchors placed at locations deemed less beneficial are discussed, showing the advantage of our proposed approach.

Experimental Design For Sampling Multiple Offspring Per Litter To Evaluate Reproductive Developmental Effects Following In Utero Through Lactational Chemical Exposure In Sprague

✦Zhenxu J Ma, Battelle, 501 King Ave., Columbus, OH 43201, *maj@battelle.org*; Paul I Feder, Battelle; Don R Bergfelt, EPA; Ralph L Cooper, EPA; David P Houchens, Battelle

Key Words: sampling design, endocrine disruptor, power analysis, mixture of normal and Poisson, mixture of normal and binomial, interlitter and intralitter variances

Interaction of certain chemicals with the endocrine system has been reported to result in reproductive developmental effects in humans and wildlife. Data for the present study were made available by the US Environmental Protection Agency to examine the added power obtained by increasing the number of F1 male or female offspring selected per litter and ability to detect statistically significant effects on hormone-dependent endpoints. Reproductive developmental endpoints were recorded as continuous, count and categorical measurements following in utero through lactational exposure to VIN, DBP and two chemical mixtures. Power analyses were conducted for each endpoint to evaluate the effects of sampling 1, 3 or 5 pups per litter involving 20 or 10 litters. Hierarchical distributions and Monte Carlo simulations were conducted, which incorporated between-pup and between-litter variabilities. Power analyses demonstrated that, regardless of litter size, there was a greater increase of power going from 1 to 3 pups per litter compared to 3 to 5 pups per litter. Moreover, by increasing number of pups sampled per litter, comparable results could be obtained with a reduced number of litters.

Latitudinal Profiles Of Seasonal Rainfall-Enso Association Along The Coast Of Central And South America, Using Time Lagged Three Way Contingency Tables

✦Luis S Cid, Universidad del Bio Bio, Universidad del Bio Bio, Concepcion, 01 Chile, *lcidserrano@gmail.com*; Sandra M Ramirez, Universidad Javeriana de Cali; Eric Alfaro, Universidad de Costa Rica

Key Words: El NiÒo, rainfall, lagged category, contingency tables

Our primary interest is to determine the probability of occurrence of wet or dry events based on the occurrence of an El NiÒo event and from there define strategies according its consequences on climate. We focus on the ocean/atmosphere interaction (ENSO) time series including sea surface temperature (SST), sea level pressure (SLP) and rainfall anomalies. We categorize an ENSO index (NOS) into terciles and dry/normal/wet for rainfall. The objective is to model the relationships between the El NiÒo and rainfall over the west coast of Central and South America. Data consist of time series of SST, SLP and rainfall for different latitudes along the west coast of Central and South America over 2.5x2.5 degrees grid. Non symmetrical three way contingency tables were considered, including a time lagged categorization of the predictor variable (NOS). Data will be analyzed using generalized linear statistical models, to generate a latitudinal mapping of the association between rainfall and NOS for the west coast of Central and South America, for latitudes starting at 10 oN, through 35 oS. We generate latitudinal profiles for the time lagged association between rainfall and the ENSO.

364 Business and Economic Statistics Section Speaker with Lunch (fee event)

Business and Economic Statistics Section

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

Challenges For Economic Policy: 2012 And Beyond

✦Gary H. Stern, Stern Partners LLC, Minneapolis, MN , *ghsstern@comcast.net*

The financial crisis and deep and prolonged recession of 2007-2009 provoked an unprecedented policy response from the Federal Reserve and from Congress and the Administration. As the dust is settling and the economy improving, it is timely to assess three critical issues: 1) financial reform (the Dodd-Frank legislation) and its prospects for effectively addressing Too-Big-To-Fail; 2) Federal budget deficits and fiscal policy alternatives; and 3) longer-term economic prospects for the U.S. in light of these challenges, and of prospects for productivity growth.

365 Biopharmaceutical Section P.M. Roundtable Discussion (fee event)

Biopharmaceutical Section

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

Recognizing Author Initiative

✦S. Stanley Young, National Institute of Statistical Sciences, 27709 USA, *young@niss.org*

Key Words: Multiple testing, Multiple modeling, Epidemiology

Publication in applied observational science areas, e.g. environment, medicine, nutrition, epidemiology, psychology, sociology, etc., generally requires a "statistically significant" result. Journal editors are largely using statistical significance as the price for entry to their journals. Edwards Deming might say, "Quality by inspection, rather than quality by design." Authors can be quite through (multiple testing) and creative (multiple modeling) to get statistical significance. We will discuss examples of author initiative and how inclined statisticians might effectively push back and provide some level of independent oversight.

Role of Statisticians and Programmers in Assessment of Safety

✦Vipin Arora, Director, Statistics, Global Biostatistics and Data Management

Abbott Laboratories, Dept NJ46, AP-9A-LL, 100 Abbott Park Road, Abbott Park, IL 60064 USA, *vipin.arora@abbott.com*

Key Words: Safety, Pooled analysis, Statisticians, Programmers

In Pharmaceutical industry, there is an increased awareness to monitor safety in view of changing regulatory landscape. It is critical that statisticians (and programmers) collaborate early on with Clinical and Safety teams to propose and finalize collection of safety data from an individual study and plan to pool the safety data across various indications (ie not limited to one submission or expansion of an existing label). Wherever possible, statistical assessments in terms of power and related measures should also be provided to stake holders to increase awareness of including a specific analysis at a study level versus the corresponding analysis at the pooled level. Examples of such collaboration(s) will be shared at the round table luncheon alongwith optimal ways of integrating and presenting safety data.

Data Mining In Vaccine Manufacturing: Finding Needles In Biological Haystacks

✦Nelson Lee Afanador, Merck, Sharp, & Dohme Corp., 770 Sumneytown Pike, PO Box 4, West Point, PA 19486, *nelson.afanador@merck.com*

Key Words: Data Mining, Random Forest, Recursive Partitioning, Vaccine Manufacturing

Vaccine manufacturing is the latest field in which advanced data mining methods are beginning to gain widespread acceptance among data analysts. Vaccine manufacturing itself is a complex biological process composed of hundreds of steps carried out over several months. During the manufacture of a single vaccine lot hundreds of processing variables and raw materials are monitored. The challenging aspects with respect to mining biological manufacturing process data begins with obtaining a suitable dataframe which to analyze, proceeds to inherent variation in raw material composition and processing conditions, and ends with high inherent variation in the measurement systems. As such, identifying the root cause candidates for changes in cellular growth or viral propagation is extremely challenging. Given the large numbers of available candidate variables the traditional methods of univariate statistical process control charting, analysis of variance, and least squares regression leave many questions unanswered. Methods such as Random Forest and single-tree recursive partitioning have proven to be important methods in helping drive at potential root causes for observed changes.

Small Event Rates, Big Outcome Studies

✦Jennifer E. Hamer-Maansson, AstraZeneca, 1800 Concord Pike, Wilmington, DE 19810 USA, *jennifer.hamer@astrazeneca.com*

Key Words: small event rates, outcome studies, relative risk, absolute risk, diabetes, asthma

The FDA now requires a sponsor to exclude a certain level of risk of certain safety concerns for select classes of compounds. For example, all diabetes treatments must exclude a certain level of risk of the compound versus all comparators for cardiovascular events and long acting beta agonists (LABAs) must exclude a certain level of risk for asthma exacerbations. The small event rates for certain safety events can make these outcome studies rather large. We will discuss the issues associated with these studies (sample size, non-inferiority margins, relative risk versus absolute risk, recruitment issues, cost, populations to be studied, choice of active or placebo control, etc.).

Use Of Propensity Score Analysis Method: Assumptions/Validity Testing And What To Do When They Fail

✦Terri Kang Johnson, FDA/CDRH, 10903 New Hampshire Ave. WO-66, Silver Spring, MD 20993, *terri.johnson@fda.hhs.gov*

Key Words: propensity score, historical control, non-randomized trial, validity testing

Often a randomized trial is deemed unfeasible or unethical in testing for the safety and effectiveness. In such cases, a non-randomized trial that utilizes a historical control data may be conducted when sufficient clinical knowledge and information are available. One concern that arises in using historical controls is differential characteristics that may exist between two comparison groups that yield a bias treatment effect. In observational studies, propensity score analysis has been performed as one way to adjust for potential confounding covariates between groups. Furthermore, there has been increased interest in applying propensity score methods to nonrandomized clinical studies as well. In either setting, propensity scores are used primarily to reduce bias and increase precision. This roundtable session is to discuss assumptions and validity testing associated with three most common types of propensity score analysis method (matching, stratification, and regression) and their adjustment methods, if any, when bias has not been fully corrected by the use of propensity score.

Statistical Issues In Futility Analyses

✦Xuan Liu, Merck, , *xuan.liu@merck.com*

Key Words: futility analysis, power, sample size re-estimation

Futility analysis has been used widely in randomized clinical trials. It is not uncommon that the experimental medicine fails to demonstrate benefit versus standard treatment. In these situations, it may be plausible to stop the trial early to reduce the exposure to inefficacious treat-

ment for patients from an ethical stand point and save resource for more promising compounds from a business stand point. In this session we will discuss issues related to the futility analysis including how to select the futility boundary and timing of the futility analyses, and the corresponding impact on the power, and sample size re-estimation.

366 Business and Economic Statistics Section P.M. Roundtable Discussion

Business and Economic Statistics Section

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

The Emergence Of Analytics In The World Of Business Decision Making

✦Aric LaBarr, Institute for Advanced Analytics at North Carolina State University, 920 Main Campus Drive, Venture 2 Building, Suite 530, Raleigh, NC 27606 USA, *aric_labarr@ncsu.edu*; Michael Rappa, Institute for Advanced Analytics at North Carolina State University

Key Words: Analytics, Business Statistics, Statistics Education, Decision Making

Due to our ever expanding, data-driven society, the availability and quantity of data to decision-making people in the business sector is growing rapidly. The emergence of the field of analytics has started a culture shift in how businesses view the role of the data analyst. This roundtable will discuss some of the current issues by both industry and academia to adjust to this changing culture. Topics to be discussed may include: the new role of the data analyst, the expectations of companies for business focused statisticians, academia's preparation of students for this changing field, and options for students to be prepared for this new culture of analytics.

367 ENAR P.M. Roundtable Discussion

ENAR

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

Statistics In Neuroimaging: Current Challenges And Emerging Issues

✦Hongtu Zhu, University of North Carolina Department of Biostatistics, 3101 McGavran-Greenberg, CB#7420, Chapel Hill, NC 27599 USA, *hzhu@bios.unc.edu*; ✦Martin A Lindquist, Columbia University, 1255 Amsterdam Ave, Room 1031, 10th Floor, MC 4690, New York, NY 10027, *martin@stat.columbia.edu*

Neuroimaging studies give rise to high-dimensional data with complicated temporal and spatial noise structure. The statistical analysis of the data is challenging and can range from determining the appropriate statistical method to apply, to developing unique methods geared toward the specific imaging modality. Due to the importance that statistics plays, it is crucial for more statisticians to get involved for the field to reach its full potential. However, it can be difficult for statisticians to get their methods used. The goal of this roundtable is to discuss ways of ensuring that statisticians make as big an impact as possible on this emerging field. This includes fostering research, education and the influence of statisticians on imaging sciences. This can be done through collaborations, sharing of statistical tools and the creation of a common forum for disseminating results and discussing emerging issues. We also discuss the goals of a recently formed special interest group of ASA - Statistics in Imaging (SI) - that seeks to address these questions. Participants should be ready to pose questions and discuss their personal experiences with neuroimaging data.

368 Section for Statistical Programmers and Analysts P.M. Roundtable Discussion (fee event)

Section for Statistical Programmers and Analysts

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

How to Submit SDTM Data for FDA Review

✦Wei (Lisa) Lin, Merck & Co., UG1CD-14, 351 N. Sumneytown Pike, North Wales, PA 19454, *lisa_lin@merck.com*

Key Words: SDTM, data, submission, challenge, solution

As a statistical programmer, what do we need to prepare for SDTM data submission? What is the challenge and solution?

369 Section on Bayesian Statistical Science P.M. Roundtable Discussion (fee event)

Section on Bayesian Statistical Science

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

Practical Issues in Bayesian Adaptive Designs for Early-Phase Clinical Trials

✦Peter Francis Thall, M.D. Anderson Cancer Center, Department of Biostatistics, P.O. Box 301402, Houston, TX 77230-1402 USA, *rex@mdanderson.org*

Key Words: Bayesian statistics, Clinical trials, Phase I trial, Phase II trial, Phase I/II trial

In recent years, there has been an explosion in published Bayesian methods for design and conduct of early phase clinical trials. These include methods for dose-finding, safety and futility monitoring, adaptive randomization, and dealing with multiple outcomes. Very little attention has been paid to practical issues arising in actual application and implementation of such designs. These include establishing priors, conducting computer simulations to evaluate design properties, and dealing with logistical or ethical issues in applying outcome-adaptive decision rules when accrual is either fast or slow relative to the timeframe for observing outcomes. In this roundtable, we will discuss any or all of these issues. Participants are encouraged to bring examples, additional issues, and questions from their own work.

370 Section on Government Statistics
P.M. Roundtable Discussion (fee event)

Section on Government Statistics

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

Balancing Data Quality And Data Confidentiality

✦Lawrence Cox, National Institute of Statistical Sciences, 12177 Etchison Road, Ellicott City, MD 21042, *statdisclosure@aol.com*

Key Words: statistical disclosure, data quality, confidentiality, suppression, rounding, synthetic data

We will discuss the goal of releasing high-quality disclosure-protected data from both qualitative and quantitative standpoints. Qualitatively: for a given data set (tabular or microdata), how should quality and confidentiality be defined, and evaluated? what does it mean to "balance" quality and confidentiality--is one more important than the other? are these dual objectives always in conflict? Quantitatively (you may wish to bring paper and pencil): how effective is complementary cell suppression for achieving high-quality disclosure-protected tabular data? how effective are controlled tabular adjustment and rounding, and pre-tabulation perturbation of underlying microdata? how efective is microaggregation for achieving high-quality disclosure-protected microdata? how effective is synthetic microdata? Contributions to the discussion of actual participant experiences will be welcome.

371 Section on Health Policy Statistics
P.M. Roundtable Discussion (fee event)

Section on Health Policy Statistics

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

Teaching Health Policy Statistics In The Pacific Islands

✦Mark Griffin, School of Population Health, University of Queensland, Herston, Brisbane, International 4006 Australia, *m.griffin@uq.edu.au*

Key Words: health, education, developing nation, conference, training, pacific islands

In July 2011 the ASA Friends of Australasia organised the first International Conference for Health Statistics in the Pacific Islands. To the best of our knowledge this is the first statistics conference to ever be held within the Pacific Islands in Australasia. During this session you will hear from the ICHSPI Conference Chair about his experiences in organizing this conference and his advice for statistical educators involved in training within the developing nation context. This conference was held in Suva, Fiji, and was organised in partnership with a number of other ASA groups (including the Health Policy Statistics Section and Statistics Without Borders) and a number of groups within Australasia (including the Fiji School of Medicine, the University of the South Pacific, and the Statistical Society of Australia). This roundtable will be led by Mark Griffin, Chair of the Friends of Australasia.

372 Section on Quality and Productivity
P.M. Roundtable Discussion (fee event)

Section on Quality and Productivity

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

What Is A Statistical Engineer And Do I Want To Be One?

✦Jennifer H Van Mullekom, DuPont, 5401 Jefferson Davis Highway, Richmond, VA 23234 USA, *jennifer.h.van-mullekom@usa.dupont.com*

Key Words: Statistical Engineering, Statistical Consulting, Quality, Productivity, Physical and Engineering Sciences, Financial Industry

Statistical engineering, as defined by Hoerl and Snee, has received a great deal of attention in journals and at conferences recently. So, what is statistical engineering? What are examples of this concept? Should statisticians be statistical engineers? What are the pros and cons? Is practicing statistical engineering good for my career? This roundtable luncheon will explore these questions and few more from the perspective of a statistical consultant with 13 years experience. The discussion will emphasize applications in quality, productivity, the physical and engineering sciences, and the financial industry. Come with your opinions and your case studies or attend to gain an introduction to the topic.

373 Section on Statistical Computing
P.M. Roundtable Discussion (fee event)

Section on Statistical Computing

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

Comparing Distance Learning Platforms For Statistics Courses

✦Seth Hirshorn, University of Michigan-Dearborn, 1310 SSB, DEARBORN, MI 48326, *shirsh@umich.edu*

Key Words: Distance Learning, Statistics Courses, Blackboard, C-tools, Software Platforms

The workshop will focus on experiences using a variety of online platforms instructing Statistics Courses. The discussion will be focused on the results of an instructor survey at the University of Michigan to assess experiences using c-tools, VLT, Blackboard and other platforms. Some of the questions that will be addressed are: Reactions to a recent Department of Education study of online post secondary instruction will be discussed. The study indicated: Providing further evidence of the tremendous opportunity to use technology to improve teaching and learning, the U.S. Department of Education today released an analysis of controlled studies comparing online and face-to-face instruction. The analysis also showed that the instruction conducted wholly on line was more effective in improving student achievement than the purely face to face instruction. In addition, the report noted that the blended conditions often included additional learning time and instructional elements not received by students in control conditions. What are the special requirements of Statistics courses and the concomitant capabilities (or lack thereof) of the instructional platforms used.

R, Finance, And Statistical Computing

✦Paul R Teetor, Self-employed, 1340 Pleasant Dr, Elgin, IL 60123, *paulteetor@yahoo.com*

Key Words: R, finance, statistical computing

In finance, R has quickly eclipsed other tools for statistical computing. R-based applications are moving out of the research departments and onto the trade desks. R packages related to finance are growing in size and sophistication, and the annual R/Finance Conference in Chicago is drawing a large, nation-wide audience from academia and industry. What challenges does the R finance community face now? What's on their leading edge, both in research and practice?

374 Section on Statistical Consulting P.M. Roundtable Discussion (fee event)

Section on Statistical Consulting

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

Using Email Newsletters, Webinars, Blogs, And Social Media To Promote Your Consulting Career

✦Stephen Simon, P Mean Conulting, 14814 Granada Court, Leawood, KS 66224 USA, *mail@pmean.com*

Key Words: Facebook, LinkedIn, Twitter, advertising, promotion

There are a wide range of resources on the Internet that you can use to advertise your services and promote your career as a statistical consultant. This roundtable session will list some of the resources that I have used (email newsletters, webinars, blogs, Facebook, LinkedIn, and Twitter). None of these resources are expensive; most, in fact, are free. They do, of course, represent a significant investment of your time. Used properly, they can greatly raise your profile in the research community. There are informal norms for conduct with each of these resources, however, that you should be respect to avoid negative publicity. Participants will be encouraged to share their experiences with these and other Internet resources.

375 Section on Statistical Education P.M. Roundtable Discussion (fee event)

Section on Statistical Education

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

Using Class Activities to Teach Statistics Appreciation Courses

✦Jamis Perrett, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, *jamis@stat.tamu.edu*

Key Words: Activities, Active Learning, Statistics Appreciation, Liberal Arts, Introductory Statistics, Statistics Education

Class time for statistics appreciation courses can be best spent using activities to learn statistical concepts and to get students engaged in actually doing statistics. The vast array of tools in the newer course management systems makes it fairly easy for students to spend time outside of the scheduled class time for textbook-type learning. This way, students are actively learning and participating in class, and class time is more productive. Students leave feeling like they have experienced statistics rather than having been subjected to another math class. This apporach is especially effective with students who come from liberal arts backgrounds, who don't need to learn all the math of an introductory statistics course as would say students from business or engineering majors. The roundtable will include discussion on the concept as well as details on the implementation.

Distance Learning Technologies in Introductory Statistics Courses

✦Jose-Miguel Yamal, University of Texas School of Public Health, 1200 Herman Pressler, RAS W928, Houston, TX 77030, *Jose-Miguel.Yamal@uth.tmc.edu*

Key Words: education, ITV, technology, distance learning

Teaching introductory statistics courses presents a challenge to students with little mathematical background. This is an even greater challenge when the students are at remote locations. Interactive television (ITV) has been used to help build productive learning environment. We will discuss some of the challenges and advantages of using technologies for distance learning and share ideas of how to teach effectively.

Teaching Statistics to Culturally and Linguistically Diverse Students

✦Lawrence M Lesser, The University of Texas at El Paso, 500 W. University Avenue, Department of Mathematical Sciences, El Paso, TX 79968-0514, *Lesser@utep.edu*

Key Words: language, diversity, education, culture, equity

Statistics faculty are increasingly likely to face classes that are more diverse not only in content background but also in demographics such as students' home culture and language. This interactive, participatory roundtable explores how these challenges can be turned into equitable opportunities for all students to learn. Lesser (an ASA Section officer as well as a founding editor of Teaching for Excellence and Equity in Mathematics) draws from his ongoing scholarship (e.g., the leadoff paper of the November 2009 Statistics Education Research Journal and an invited refereed paper in the 2010 ICOTS Proceedings; both papers are at http://www.stat.auckland.ac.nz/~iase/publications.php) and experiences teaching statistics to English learners at a research university with a 21st-century student demographic. We'll learn from each other's resources, strategies, successes, and lessons learned.

Teaching Statistics With R

✦Randall Pruim, Calvin College, Grand Rapids, MI 49546, *rpruim@calvin.edu*

Key Words: education, computation, R

R is a powerful platform for statistical analysis and data visualization, but is it a good tool for teaching statistics to undergraduates? In this round-table we'll discuss * types of students and course where R is more or less appropriate, and possible alternatives; * advantages of using R; * challenges in using R with beginners, and ways to reduce or overcome them; * resources for teaching statistics with R, including

some new R packages by the Project MOSAIC team, and text books that work well with R; * R computing environments including R Studio and Rcmdr; * what instructors new to R need to know, and how to learn it. Come prepared to ask questions and to share your experiences using R and other products in statistics courses at all levels.

376 Section on Survey Research Methods P.M. Roundtable Discussion (fee event)

Section on Survey Research Methods

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

Research at the U.S. Census Bureau

✦Daniel Weinberg, US Census Bureau, ADRM - Room 8H116C, Washington, DC 20233, *daniel.h.weinberg@census.gov*

Key Words: Survey methodology, Record linkage, Confidentiality, Economic surveys, Adminstrative records, Federal statistical system

Recognizing the importance of research to find improved ways of conducting censuses and surveys, the U.S. Census Bureau has set up a new Research and Methodology Directorate. Roderick Little, Associate Director for Research and Methodology at the Census Bureau, will describe the organization of the directorate, some of the key statistical problems to be addressed, and exciting employment opportunities arising from this initiative.

Survey Quality Indicator Measures: Response Rates and Alternatives

✦Donsig Jang, Mathematica Policy Research, Inc., 600 Maryland Ave., SW, Suite 550, Washington, DC 20024, *djang@mathematica-mpr.com*

Key Words: Paradata, Survey quality measure, R-indicators, Nonresponse bias, Auxiliary variables

Response rate is often used as an indication of measuring the quality of the survey response. However, it only tells one side of the survey story; the other side about the association between respondents and nonrespondents are unknown. Researchers continue to seek the tools to assess and compare the quality of the response to different surveys. For example, introduced by Schouten et al. (2009), R-indicators are used to measure how well a respondent set represents the sample or population from which it was drawn. This measure may be a better indicator of survey nonresponse bias than response rates for survey outcomes closely related to auxiliary variables used for R-indicator calculation. In this roundtable, we will lead a discussion of the use of alternatives to response rates in measuring survey quality.

377 Social Statistics Section P.M. Roundtable Discussion (fee event)

Social Statistics Section

Tuesday, August 2, 12:30 p.m.–1:50 p.m.

American Community Survey Data Products - Past And Future

✦Trent Alexander, U.S. Census Bureau, 4600 Silver Hill Rd., Room 3K071, American Community Survey Office, Census Bureau, Washington, DC 20233 USA, *j.trent.alexander@census.gov*

Key Words: U.S. Census Bureau, American Community Survey, Social statistics, Economic statistics

Since there was no long-form census in 2010, American Community Survey data have become the most relevant and timely source for a broad range of social and economic statistics, especially for small geographic areas. In December 2010, the Census Bureau released the first-ever 5-year estimates from the ACS. This roundtable will discuss the potential uses of these data, as well as the ways in which the Census Bureau could change or improve ACS data products in the next few years.

378 Recent Advances in Statistical Genomics ■●

Biometrics Section, ENAR, International Indian Statistical Association, Section on Statistics in Epidemiology

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Using Phylogenetic Relationships to Infer Gene Functional Groups

✦Jun Liu, Harvard University, Department of Statistics, Cambridge, MA 02138, *jliu@stat.harvard.edu*; Yang Li, Shanghai JiaoTong University; Roee Gutman, Harvard University

Key Words: Phylogeny, Markov chain Monte Carlo

When a set of genes (proteins) appear and disappear simultaneously from many different species, it suggests that these genes are likely functionally related. We developed novel method to infer such functional relationship among the genes by directly modeling the phylogenetic relationships among the species. Simulation studies and experimental validations show that the method is superior to other heuristic approaches.

A Bayesian Measurement Error Model for Two-Channel Cell-Based RNAi Data with Replicates

✦I-Shou Chang, National Health Research Institutes, 350 Taiwan, *ischang@nhri.org.tw*; Chung-Hsing Chen, National Health Research Institutes; Chao Hsiung, National Health Research Institutes; King-Song Jeng, Academia Sinica

Key Words: Bayesian models, HCV replication, High-throughput screening, Multiple hypothesis testing, RNA interference, Viral-host interaction

RNA interference (RNAi) is an endogenous cellular process in which small double-stranded RNAs lead to the destruction of mRNAs with complementary nucleoside sequence. With the production of RNAi libraries, large-scale RNAi screening in hu- man cells can be conducted to identify unknown genes involved in a biological pathway. One challenge in these studies is to deal with the multiple testing issue and the related false positive rate (FDR) and false negative rate (FNR). We pro-

pose a Bayesian measurement error model for the analysis of data from a two-channel RNAi high-throughput experiment with replicates, in which both the activity of a particular bio logical pathway and cell viability are monitored and the goal is to identify shRNAs that affect the pathway activity without affecting cell activity. Simulation studies indicate the excellent numerical performance of this method and provide insight into the effects of prior distributions and the number of replicates on FDR and FNR. This method is illustrated in analyzing the data from a RNAi high-throughput screening that searches for cellular factors affecting HCV replication without affecting cell viability

Another Look at Inference on Ranked Observations, with Application to Genomics

✦Fred Andrew Wright, Univ North Carolina, 4115B McGavran-Greenberg, CB#7420 Univ North Carolina, Chapel Hill, NC 27599, *fwright@bios.unc.edu*; Yihui Zhou, Univ North Carolina

Key Words: statistical genetics, winner's curse, estimation

We explore a new approach to estimation of effect sizes among ranked hypotheses, motivated by the need to perform estimation in the context of multiple testing. The need to reduce bias for ranked observations is exemplified in various approaches to "winner's curse" correction in genetic association analysis, but appear in a wider variety of problems. We propose an approach to reduce ranking bias via a combination of the jackknife and computation of leave-one-out pseudo-values. Our approach dramatically reduces the effects of ranking bias, but computation of appropriate standard errors for ranked hypotheses remains challenging. We describe a proposed resampling approach to obtain the standard errors.

379 Latest developments on Analysis of Missing Data ●

Section on Health Policy Statistics, ENAR, Section on Government Statistics, Section on Survey Research Methods, Social Statistics Section

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

A Functional Method for Longitudinal Data with Missing Responses and Covariate Measurement Error

✦Grace Y Yi, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, N2L 3G1 Canada, *yyi@uwaterloo.ca*; Yanyuan Ma, Department of Statistics, Texas A&M University; Raymond James Carroll, Texas A&M University

Key Words: Functional measurement error modeling, Generalized method of moments, Inverse probability weighting, Longitudinal data, Missing response

Covariate measurement error and missing responses are two typical features in longitudinal data analysis. There has been extensive research on either covariate measurement error or missing responses, but relatively little work has been done to address both characteristics simultaneously. In this talk, we propose a simple method for the marginal analysis of longitudinal data with time-varying covariates, some of which are measured with error, while the response is subject to missingness. The proposed method has a number of appealing properties: assumptions on the model are minimal, including no assumptions about the distribution of the mismeasured covariate; implementation is quite straightforward; and the applicability of the proposed method is broad. We provide both theoretical justification and numerical results of our method.

Generating Multiple Imputations from Multiple Models to Incorporate Model Uncertainty in Nonignorable Missing Data Problems

✦Ofer Harel, University of Connecticut, 215 Glenbrook Road U-4120, Storrs, CT 06269, *oharel@stat.uconn.edu*; Juned Siddique, Northwestern University

Key Words: multiple imputation, misssing data, nonignorable models, model uncertainty

We present a framework for generating multiple imputations when the missing data are assumed to be non-ignorably missing. Imputations are generated from more than one model in order to incorporate uncertainty regarding what is the "correct" imputation model. Parameter estimates based on the different imputation models are combined using the rules of nested multiple imputation. Through the use of simulation, we investigate the impact of imputation model uncertainty on post-imputation inferences and show that incorporating model uncertainty can improve the coverage of parameter estimates. We also apply our method to a longitudinal depression treatment study. Our method provides a simple approach for formalizing subjective notions regarding non-response so that they can be easily stated, communicated, and compared.

Empirical Likelihood-Based Method Using Calibration for Longitudinal Data with Drop-Out

✦Xiao-Hua Andrew Zhou, Department of Biostatistics, University of Washington, 1705 N.E. Pacific Street, Seattle, WA 98195, *azhou@uw.edu*; Baojiang Chen, Department of Biostatistics

Key Words: missing-data, calibration, empirical likelihood, longitudinal data

In longitdinal studies, interest often lies in estimation of the population-level relationship between the explanatory variables and dependent variables. In this talk we propose an empirical likelihood-based method to incorporate population level information in a longitudinal study with drop-out. The population-level information is incorporated via constraints on functions of the parameters, and non-random drop-out bias is corrected by using a weighted generalized estimating equations method. We provide a three-step estimation procedure that makes computation easier. Several methods that are often used in practice are compared in simulation studies, which demonstrate that our proposed method can correct the non-random drop-out bias and increase the estimation efficiency, especially for small sample size or when the missing proportion is high. Also, the proposed method is robust to misspecification of the working correlation matrix or the missing data model under the missing at random mechanism. Finally, we apply this method to an Alzheimer's disease study. This is a joint work with Dr. Baojing Chen and Gary Chan.

Bayesian Modeling and Inference for Data with Informative Treatment Switching or Dropout

✦Ming-Hui Chen, University of Connecticut, Department of Statistics, 215 Glenbrook Road, U-4120, Storrs, CT 06269, *mhchen@stat.uconn.edu*; Qingxia Chen, Vanderbilt University; David Ohlssen, Novartis Pharmaceuticals Corporation; Joseph G. Ibrahim, University of North Carolina

Key Words: Intermittent missingness, Markov chain Monte Carlo, Missing at random, Multivariate logistic regression, Multivariate mixed-effects model, Non-ignorable

In randomized clinical trials, it is common that patients may stop taking their assigned treatments and then start the standard treatment or completely dropout from the study. In addition, patients may miss scheduled visits even during the study, leading to intermittent missingness. In this paper, we develop a novel Bayesian method for jointly modeling longitudinal treatment measurements under various dropout scenarios. Specifically, we propose a multivariate normal mixed-effects model for repeated measurements from the assigned treatments and the standard treatment, a multivariate logistic regression model for those stopping the assigned treatments, logistic regression models for those starting a standard treatment off protocol, and a conditional multivariate logistic regression model for completely withdrawing from the study. We assume that withdrawing from the study is non-ignorable but intermittent missingness is assumed to be at random. Various properties of the proposed model are examined. An efficient Markov chain Monte Carlo sampling algorithm is developed. A real data set from a clinical trial is analyzed in detail via the proposed method.

380 Nonparametric Modeling and Analysis

IMS, International Chinese Statistical Association, International Indian Statistical Assoc., Reps. for Young Statisticians, Section on Nonparametric Statistics, SSC

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Nonparametric Regression and the Secondary Analysis of Case-Control Studies

✦Raymond James Carroll, Texas A&M University, 3143 TAMU, Department of Statistics, College Station, TX 77843-3143, *carroll@stat.tamu.edu*; Jiawei Wei, Texas A&M University

Key Words: Case-control studies, Nonparametric regression, Semiparametric regression, Biased sampling

Case-control studies collect information on case-control status and covariates. Increasingly, investigators have wished to exploit case-control data sets to explore relationships among the covariates themselves. Thus, case-control status is D, and covariates are (Y,X), and the goal is to make inference about the relationship of Y and X. One cannot simply use the (Y,X) data directly because of the biased sampling caused by the case-control design, although with rare diseases regression of Y on X among the controls is essentially unbiased. The goal though is not to throw away the case data, but to exploit it for more efficient inference. Efficient semiparametric approaches exist if there is a parametric model relating Y to X, but these approaches can lead to considerable bias if the model is incorrectly specified. We develop parametric and nonparametric methods that are robust to model misspecification and have much greater efficiency than analysis of the control data only.

A Unified Approach for Support Vector Regression Under Right Censoring

✦Yair Goldberg, UNC-CH, Department of Biostatistics,, UNC-CH, Chapel Hill, NC 27599, *ygoldber@bios.unc.edu*; Michael R Kosorok, UNC-CH

Key Words: survival analysis, support vecotor machine, generalization error

We develop a unified approach for support vector machines for classification and regression in which the outcomes are functions of the survival times subject to right censoring. We present a novel support-vector regression algorithm that is adjusted for censored data. We provide finite sample bounds on the generalization error the algorithm. We apply the general methodology for estimation of the (truncated) mean, median, quantiles, and for classification problems.

Variable Selection in High-Dimensional Varying Coefficient Models

Lan Xue, Oregon State University; ✦Annie Qu, University of Illinois at Urbana-Champaign, 101 Illini Hall, 725 S Wright St., Champaign, IL 61820 USA, *anniequ@illinois.edu*

Key Words: difference convex programming, L_0-regularization, large-p small-n, nonparametric function, oracle property

The varying coefficient model is flexible and powerful for modeling the dynamic changes of regression coefficients. Here the response variables depend on covariates through linear regression, but the regression coefficients can vary and are modeled as a nonparametric function of other predictors. It is important to identify significant covariates associated with response variables, especially for high dimension setting where the number of covariates can be larger than the sample size, but the number of signal terms is relatively smaller than the sample size. We consider model selection in such setting and adopt difference convex programming to approximate the L0 penalty, and investigate global optimality properties of the varying coefficient estimator. The challenge of the variable selection problem here is that the dimension of the nonparametric form for the varying coefficient modeling could be infinite, in addition to dealing with the high-dimensional linear covariates. We show that the proposed varying coefficient estimator is consistent, enjoys the oracle property and achieves an optimal convergence rate for the non-zero nonparametric components for high-dimensional data. Our

Is Sparseness the Answer to Model Selection as Well as Prediction?

✦peter j Bickel, UC Berkeley, 367 Evans hall, UC Berkeley, Berkeley, CA 94720, *bickel@stat.berkeley.edu*; Ya'akov Ritov, Hebrew University ,Jerusalem

Key Words: nonparametric regression, model selection, Lasso

There has been considerable emphasis on thresholding and the LASSO as computable methods leading to sharp Oracle bounds and optimal prediction rates with high dimensional covariates in regression and more

generally(Buhlmann,Meinshausen,Yu,Candes,Tao,Ritov,Tsybakovand many others). These results are predicated on the existence of unique sparsest representations of the regression. There has been considerable extension of these inquiries into identifying the variables entering into such representations of the regression.,or equivalently model building under the same conditions as those used in prediction(Wainwright and others). We argue that in many situations, for instance biological network identification ,there are many optimal predictors involving quite different combinations of variables. We present serious failures of the Lasso in such situations and discuss some alternative approaches

381 Analysis of Complex Time Series Data ■●

IMS, International Chinese Statistical Association, International Indian Statistical Association

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Covariance Matrix Estimation in Time Series

✦Wei Wu, University of Chicago, 5734 S University Ave, Chicago, IL 60637 USA, *wbwu@galton.uchicago.edu*

Key Words: Covariance matrix estimation

I will discuss estimation of covariance matrices of stationary processes. Under a short-range dependence condition for a wide class of nonlinear processes, I will show that the banded covariance matrix estimates converge in operator norm to the true covariance matrix with explicit rates of convergence. I will also consider the consistency of the estimate of the inverse covariance matrix. These results are applied to a prediction problem, and error bounds for the finite predictor coefficients are obtained. The work is joint with Mohsen Pourahmadi of TAMU.

Dynamic Modeling and Prediction of Risk Neutral Densities

✦Rong Chen, Rutgers University, Department of Statistics, Piscataway, NJ 08854, *rongchen@stat.rutgers.edu*

Risk neutral density is extensively used in option pricing and risk management in finance. It is often implied using observed option prices through a complex nonlinear relationship. In this study, we model the dynamic structure of risk neutral density through time, investigate modeling approach, estimation method and prediction performances. State space models, Kalman filter and sequential Monte Carlo methods are used. Simulation and real data examples are presented.

Locally Stationary Approaches to Complex Time Series

✦Guy Nason, University of Bristol, School of Mathematics, University of Bristol, BRISTOL, International bs8 1tw England, *g.p.nason@bristol.ac.uk*

Key Words: locally stationary, time series, aliasing

Many real life time series are not stationary and can often be successfully modeled using locally stationary series. We consider the case of locally stationary series subjected to aliasing and demonstrate what might be done to mitigate the effect of aliasing on the series, to gain information on the aliases and the variation in the series that is obscured by aliasing

382 The Human Cultural and Social Landscape ■●

Section on Statistics in Defense and National Security, International Chinese Statistical Association, Section on Physical and Engineering Sciences, Section on Risk Analysis, Social Statistics Section

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Understanding Afghanistan

✦Yasmin H Said, George Mason University, MS 6A2, 4400 University Drive, Fairfax, VA 22030 USA, *ysaid99@hotmail.com*

Key Words: Social networks, agent-based models, cultural landscape

The early history of Afghanistan is written in terms of rival tribal leaders of scattered and ethically diverse populations. The country itself is a barren territory which does not produce enough revenue for a central government to rule it. Thus, a regional tribal form of government naturally developed in order to eke out a subsistence living for the members of the tribes. Modern Afghanistan was created as a buffer state by the British between colonial India and imperial Russia. Afghanistan has only been organized with a central government by virtue of intervention of outside powers; first Britain, later the Soviet Union, and, most recently, Pakistan, who created the Taliban. Without external interventions, the country tends to revert to tribal/patrilineal regional rulers. This talk will discuss the role of the history, the Pashtuns, and the culture of Pashtunwali in shaping modern Afghanistan. This talk is intended to provide the logic underlying the need for a quantitative assessment of social and cultural issues associated with Afghanistan and similar tribal-oriented societies.

Validating a Model of Afghan Drug Industry: Effects of Corruption on the Effectiveness of Counternarcotic Policies

✦Armando Geller, George Mason University, , *ageller1@gmu.edu*; Maciej M. Latek, George Mason University; Seyed M. Mussavi Rizi, George Mason University

Key Words: Afghanistan, Multiagent simulation, Validation

We report the findings of a multiagent model in which we examine the effectiveness of counternarcotic policies that suppress drug production and drug exports. To do so, we first study how opportunities for daily corruption provide incentives or disincentives to farmers and traders to comply with government regulations. We then show how the outcomes of such micro decisions percolate through alternative governance institutions, local strongmen and the Taliban. Finally, we discuss the cognitive and computational foundations of decision making for self-interested economic agents, formal government institutions and informal government structures that enable them to operate plausibly in a conflict environment where corruption is possible. Our approach is focused on implementing bounded rationality while achieving culture and context plausibility and narrative face validity.

Inferring Social Network Structure from Incident Size Distribution in Iraq

✦Tim Gulden, George Mason University, 4400 University Drive, MS 6B2, Fairfax, VA 22030, *tgulden@gmu.edu*

Key Words: Social networks, power-laws, conflict, counterinsurgency, violence

The violence in Iraq between 2003 and the present has been multi-faceted and extremely hard to characterize in terms of its motivations, participants and even overall scale. This work identifies a remarkably stable truncated power-law pattern in the size distribution of violent incidents in the Iraq Body Count (IBC) database and seeks to explain it in terms of social networks and the US role in breaking up major violent groups. These results are compared to data from the leaked SIGACTS database thus generating additional hypotheses about how much of the observed effect is due to the dynamics of the conflict as opposed to the nature of the reporting.

383 Using paradata for field management, quality assurance, and statistical control ■●

Section on Survey Research Methods, Section on Government Statistics, Section on Government Statistics, Social Statistics Section, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

New Indicators and Measures to Assess Interviewer Performance in CATI Surveys

✦Francois Laflamme, Statistics Canada, Tunneys pasture, Jean Talon building, 6th floor C2, Ottawa, ON K1A 0T6 Canada, *francois.laflamme@statcan.gc.ca*

Key Words: paradata, productivity, Call center, Call scheduler

Interviewer Performance, defined in this paper as the ability of an interviewer to contact and convince respondents, is generally assessed by survey research call centres in using descriptive measures such as the number of completed interviews, the number of completed interviews per hour, etc. Other more comprehensive performance indicators such as the cooperation rate at first contact and Net Contribution to Performance Index have been developed over the past few years. However many factors might impact interviewers' performance in a centralized call centre environment. In addition to the interviewer's characteristics and environmental factors, the type and portfolio of cases called, the effort already put into these cases, the time the call is made and the general productivity of the survey at the moment at which the call is made are some of these potential influencing factors. This paper proposes a new objective interviewer's performance measure that takes into account the complexity of survey data collection process as well as new factors to consider when assessing refusal conversion and tracing data collection activities.

Comparing CAPI Trace File Data and Quality Control Reinterview Data as Methods of Maintaining Data Quality

✦Matt Jans, US Census Bureau, 8520 Greenwood Ave Apt 2, Takoma Park, MD 20912 US, *matthew.e.jans@census.gov*; Robyn Sirkis, US Census Bureau; Christina Schultheis, US Census Bureau; Renee M Gindi, National Center for Health Statistics; Jim Dahlhamer, National Center for Health Statistics

Key Words: paradata, CAPI, data quality, interviewers, NHIS, Census Bureau

Quality control is paramount to surveys conducted by the US Census Bureau. Two methods currently used to measure data quality are the Quality Control (QC) Reinterview and the Performance and Data Analysis (PANDA) system. Reinterview is a verification interview with respondents that asks questions about the interview experience to detect falsification. PANDA uses CAPI trace files, data files, and other case information (e.g., interview date and time) as indicators of cases or interviewers that might risk overall data quality (e.g,. overnight interviews). This paper explores whether these systems capture the same cases and interviewers. Both Reinterview and PANDA are used on the National Health Interview Survey, a nationwide face-to-face CAPI survey sponsored by the National Center for Health Statistics. Using data quality results from 2008-2009, we analyze which interviewers are identified by Reinterview, by PANDA, or by both. We use a multinomial logistic regression predicting identification by Reinterview, PANDA, both, or neither to see if any sample or caseload factors predict identification. We explore random interviewer and region effects on identification.

Using Statistical Process Control to Understand Variation in Computer-Assisted Personal Interviewing Data

✦Robyn Sirkis, US Census Bureau, , *robyn.b.sirkis@census.gov*; Matt Jans, US Census Bureau; Jim Dahlhamer, National Center for Health Statistics; Renee M Gindi, National Center for Health Statistics; Benjamin Duffey, University of Michigan

Key Words: Paradata, Statistical Process Control, NHIS

This paper discusses research using statistical process control with paradata obtained during data collection for the National Health Interview Survey (NHIS). Statistical process control (SPC) involves using statistical techniques to measure and analyze variation in operational processes. The goal is to not simply monitor, but to improve the quality of the process over time. For this paper, we group interviewers into clusters based on housing unit and demographic characteristics and produce control charts which examine the variation of the process over time for each cluster. We compare the means of interviewers within each cluster to determine if they are significantly different from the overall mean of the process, and examine some of the potential causes of process variation using selected control charts and SPC techniques. We address advanced SPC techniques such as multivariate charting. Indicators of data quality used in the paper are item nonresponse and interview duration. The charts are intended to demonstrate how survey managers can use paradata to monitor the data collection process using SPC principles and techniques.

An Attempt to Reduce Survey Costs via Logistic Regression and Paradata

✦Frost A. Hubbard, Institute for Social Research - University of Michigan, PO Box 1248, G386 Perry, Ann Arbor, MI 48106-1248, *fhubbard@isr.umich.edu*; James R. Wagner, Institute for Social Research - University of Michigan

Key Words: Paradata, Health and Retirement Study, Responsive Design, Institute for Social Research

Creating new methods for reducing survey costs and errors is an increasingly important issue as survey response rates continue to decline. To reduce survey costs on the Health and Retirement Study, we developed a logistic regression model that predicts the likelihood of a sampled address completing the screening interview. The screening completion propensity model will be used in two ways. First, the propensity scores for all cases will be calculated each day and matched with the cases the interviewer attempted the previous night. With this we will offer suggestions to make their contact attempts more efficient. Second, when selecting a two-phase sample, non finalized cases with higher propensities to complete the screening interview will be oversampled to increase the efficiency of the second phase sample. To determine if the screening propensity model helped reduce costs, we will compare cost metrics from before the use of the model and after. Finally, we will discuss how to further this process so that it also helps reduce the potential for nonresponse bias by using predicted values of key statistics of interest.

Analyzing Interviewer Call Record Data Using a Multilevel Multinomial Modeling Approach to Understand the Process Leading to Cooperation or Refusal

✦Julia D'Arrigo, University of Southampton, Highfiled Campus, Building 39, Southampton, International SO171BJ United Kingdom, *j.darrigo@southampton.ac.uk*; Gabriele B. Durrant, University of Southampton; Fiona Steele, University of Bristol

Key Words: paradata, interviewer call data, multilevel multinomial logistic regression

In interview-based household surveys, effective interviewer calling behaviors are critical in achieving cooperation and reducing the likelihood of refusal. This paper aims to analyze best times to gain cooperation from sample members in a range of face-to-face surveys. Of particular interest is to what extend the best times may vary with the type of household, such as a single household or a household with children. We use an unusually rich dataset, the UK Census Link Study, which combines paradata from six UK surveys, including detailed call record data, interviewer observations about the household and information about the interviewer-household interaction, which was linked to information about the household from the UK Census. The data have a multilevel structure with households nested within a cross-classification of interviewers and areas. A multilevel multinomial logistic regression approach which jointly models the different types of outcomes at each call is used to predict the likelihood of interview or refusal. The findings may have important implications for survey practice for determining best times of interviewer calls.

384 Statistics in Biosciences: Innovative Methods in Comparative Effectiveness Research ■

International Chinese Statistical Association, International Indian Statistical Association, Section on Statistics in Epidemiology, SSC

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Using Multiple Control Groups to Address Unobserved Biases in Comparative Effectiveness Research

✦Sharon-Lise Theresa Normand, Harvard Medical School, Department of Health Care Policy, 180 Longwood Avenue, Boston, MA 02115 USA, *sharon@hcp.med.harvard.edu*; Frank Yoon, Harvard Medical School; Haiden Huskamp, Harvard Medical School; Alisa Busch, McLean Hospital

Key Words: causal inference, quasi-experiments, fine balancing, propensity scores, testing in order

Studies of policy interventions typically do not involve randomization. Adjustments, such as matching, can remove the bias due to observed covariates, but residual confounding remains a concern. In this paper we introduce two analytical strategies to bolster inferences using observational data. First, we use a design with multiple comparison groups by identifying how the study groups may differ and selecting a second comparison group on this source of difference. Second, we match subjects using a strategy that finely balances the distributions of key categorical covariates and stochastically balance on other covariates. An observational study of the effect of parity on the severely ill subjects enrolled in the Federal Employees Health Benefits (FEHB) Program illustrates our methods. Comparison subjects, who did not receive parity benefits, were enrolled in matched, private plans on the basis of location and plan type. We use a second comparison grroup of FEHB subjects who were followed before parity was enacted in 2001; they are no different than those who received parity benefits, because they are from the same population.

Estimating Decision-Relevant Comparative Effects Using Instrumental Variables

✦Anirban Basu, University of Washington, Seattle, , *basua@uw.edu*

Key Words: instrumental variables, heterogeneity, local effects, local IV methods, prostate cancer

Instrumental variables methods (IV) are widely used in the health services and biostatistics literature to adjust for hidden selection biases in observational studies when estimating treatment effects. When treatment effects are heterogeneous in the population and when when individuals' self-selected choices of treatments are correlated with expected idiosyncratic gains or losses from treatments, interpretation of IV results becomes challenging. I present an overview of the challenges that arise with IV estimators in the presence of effect heterogeneity and how we can overcome these challenges. I compare conventional IV analysis with alternative approaches that use IVs to estimate treatment effects in models with response heterogeneity and self-selection. Using SEER-Medicare linked data, I apply the method of local instrumental variables to estimate the Average Treatment Effect (ATE) and the Effect on the Treated (TT) on 5-year direct costs of surgery, radiation therapy

and watchful waiting among male Medicare beneficiaries (aged 66 or older) with newly diagnosed prostate cancer. Our results reveal some of the advantages and limitations of conventional and alternative IV methods.

The Parametric G-Formula for Comparative Effectiveness Research

✦Miguel A. Hern·n, Harvard School of Public Health, 677 Huntington Avenue, Kresge, 820, Boston, MA 02115, *miguel_hernan@post.harvard.edu*

Key Words: g-formula, observational studies, comparative effectiveness research

The increasing availability of complex longitudinal databases presents both challenges and opportunities for comparative effectiveness research. A key challenge is the implementation of statistical methods that can (1) compare clinically relevant dynamic regimes, and (2) appropriately adjust for measured time-varying confounding and selection bias. Semiparametric methods such as inverse probability weighting of marginal structural models and g-estimation of nested structural models can be used to achieve these goals. However, these methods do not easily accommodate some classes of dynamic regimes, and semiparametric estimates may have a large variance when obtained from many currently available databases. A fully-parametric alternative to these methods is the parametric g-formula. Though first proposed by Robins in 1986, the g-formula has been little used in practice, and thus its relevance for comparative effectiveness research remains largely unexplored. This presentation discusses recent refinements, software developments, and practical implementations of the parametric g-formula.

385 The World of Statistical Analysis Professionals ■

Section for Statistical Programmers and Analysts, International Chinese Statistical Association, Section on Statistical Computing, Section on Statistical Consulting

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Reading the Pulse on Statistical Analysis in Nursing Research

✦Matthew J Hayat, Johns Hopkins University, 525 N. Wolfe St, Office 532, Baltimore, MD 21205 USA, *mhayat2@son.jhmi.edu*

Key Words: nursing research, statistical analysis, biostatistics consulting

Statistical analysis is a topic of interest, relevance, and often discomfort, for many nursing researchers. In this talk, we will discuss the use of statistics in the nursing field, including an overview of statistics education for nurses, common study designs used in nursing research studies, and the types of statistical methods used and published in the nursing literature. We will describe the process of developing a new biostatistics consulting service center in a School of Nursing and discuss experiences had and lessons learned.

Predictive Modeling in Home Health Care

✦Iordan Slavov, Visiting Nurse Service of New York, 1250 Broadway, 20th floor, Center for Home Care Policy and Research, New York, NY 10001, *iordan.slavov@vnsny.org*

There were significant changes lately in the ways available data are used in Home Health Care. Big companies such as the Visiting Nurse Service of NY (VNSNY) seek full utilization of their data through predictive modeling and other advanced statistical analyses. I'll discuss types of predictive models (prognostic vs. diagnostic) being built to support operations and the underlying methodology (regressions, classifications, clusters, survival analysis, etc.). Special attention will be given to service utilization models. The corresponding software skills and ways to access data and statistical knowledge will be examined. The importance given to research and analytics at VNSNY led to the formation of the Center for Home Care Policy & Research more than 15 years ago. Light will be shed on the center's position in the field of Home Care and my role as Senior Statistical Analyst there in order to benefit future professional trajectories.

The Role of Statisticians in Financial Industry Decisionmaking

✦Antonello Loddo, Capital One, 3200 Shandwick pl Apt 203, Fairfax, VA 22031 UNITED STATES~1, *antolod@gmail.com*

Key Words: financial services, analyst, business decisions, professional skills

The role of statistical analysis in the financial service industry has changed dramatically due to ever increasing data availability. As a consequence, countless opportunities for data professionals, both inside financial corporations and in analytics companies servicing the sector, have been created. Capital One was founded on the belief that the power of information, technology and testing could be combined to bring highly customized financial products directly to consumers. Statistical analysts play important roles in influencing its decision-making. We will examine how statistics help financial companies like Capital One to make better decisions and describe the skills needed to excel in the job, which skills are usually learned in school and which need to be developed as a professional. We will share opinions of professionals statisticians and their clients, and examine the career path of data professionals in the financial industry.

Statistical Analysis Abstract

✦Diahn Allen, DLA Consulting LLC, 801 S. Wells, #510, Chicago, IL 60607 USA, *diahn_allen@yahoo.com*

Objective: To illustrate how to carve a unique career path for analytical/technical professionals. My focus will be to discuss my academic background, my current and previous positions in the financial and healthcare sectors and the major functions of those positions. In particular, I will discuss what types of analyses were done, what skills were needed to be successful, what skills I acquired from those positions that are most useful elsewhere, and how those positions led to where I am today. I will also discuss future roles I anticipate.

Inside the Life of a Statistician at the National Agricultural Statistics Service

✦Matt Gregg, National Agricultural Statistics Service, 1400 Independence Ave, SW, Washington, DC 20250, *Matthew_Gregg@nass.usda.gov*

Key Words: SAS, programming, survey sampling, data analysis

The mission of the National Agricultural Statistics Service (NASS) is to provide timely, accurate, and useful statistics in service to U.S. agriculture. To meet this mission, NASS conducts hundreds of surveys every year and prepares reports covering virtually every aspect of U.S. agriculture. Statisticians at NASS are involved in all aspects of survey work including questionnaire design, frame maintenance, sample design, data collection, editing, analysis, summarization, estimation, and dissemination. During the course of a career there will be opportunities to work in all of these areas, and a person can specialize depending on his or her interests. Statistical analysis is an important part of any job at NASS. Whether performing ad hoc analysis, programming enterprise level edit, summarization, and analysis systems, or working in research and development, there is a strong need for statistical analysis professionals. Significant resources are put into developing and maintaining a high performing staff that can meet the needs of NASS now and into the future.

386 Statistical inference for functional data: theory and applications ■●

General Methodology, Section on Nonparametric Statistics

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

A Confidence Corridor for Sparse Longitudinal Data Curves

✦SHUZHUAN ZHENG, Department of Statistics and Probability, Michigan State University, A413 Wells Hall, Michigan State University, East Lansing, MI 48824, *zheng@stt.msu.edu*; Lijian Yang, Michigan State University; Wolfgang Karl Hardle, Center for Applied Statistics and Economics, Humboldt-University zu Berlin

Key Words: Longitudinal data, Confidence band, Local linear estimator, Extreme value, Double sum, Strong approximation

Longitudinal data analysis is a central piece of statistics. The data are curves and they are observed at random locations. This makes the construction of a simultaneous confidence corridor (SCC) (confidence band) for the mean function a challenging task on both the theoretical and the practical side. Here we propose a method based on local linear smoothing that is implemented in the sparse (i.e., low number of nonzero coefficients) modelling situation. An SCC is constructed based on recent results obtained in applied probability theory. The precision and performance is demonstrated in a spectrum of simultaneous and applied to growth curve data. Technically speaking, our paper intensively uses recent insights into extreme value theory that are also employed to construct a shoal of confidence intervals (SCI).

Spline Confidence Envelopes for Covariance Function in Dense Functional/Longitudinal Data

✦Guanqun Cao, Michigan State University, East Lansing, MI 48824, *cao@stt.msu.edu*; Li Wang, University of Georgia; Yehua Li, University of Georgia; Lijian Yang, Michigan State University

Key Words: B spline, confidence envelope, covariance function, functional data, Karhunen-Loeve L2 representation, longitudinal data

We consider nonparametric estimation of the covariance function for dense functional data using tensor product B-splines. The proposed estimator is computationally more efficient than the kernel-based methods. We develop both local and global asymptotic distributions for the proposed estimator, and show that our estimator is as efficient as an oracle estimator where the true mean function is known. Simultaneous confidence envelopes are developed based on asymptotic theory to quantify the variability in the covariance estimator and to make global inferences on the true covariance. Monte Carlo simulation experiments provide strong evidence that corroborates the asymptotic theory. A real data example on Tecator infrared spectroscopy data is also provided to illustrate the proposed method.

Sparse Functional Data with a Periodic Component: Methods and Applications to Psychiatric Data

✦Catherine Ann Sugar, University of California, Los Angeles, Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90095-1772, *csugar@ucla.edu*

Key Words: Functional Data Analysis, Periodic Data, Bipolar Disorder, Psychiatry

Many psychiatric illnesses have a periodic component, with patients experiencing illness "episodes" of varying length and severity, interspersed with periods of relatively good health. Even when subjects are followed longitudinally over an extended time frame the resulting data are often difficult to analyze, both because of the irregularity in the amplitude, period length and sequencing of events, and because observations tend to be sparse relative to the natural illness cycle. In this talk, we present a functional data analytic approach to such problems. Sparsity is dealt with by borrowing strength across subjects using random effects models, by jointly examining trajectories of multiple outcomes per subject, and by using pre-treatment or other supplementary data to better assess within subject periodicity. The methods will be illustrated with an application to bipolar disorder where the goal is to look at differential treatment effects on length, frequency and severity of manic and depressive episodes.

Adaptive, Robust Functional and Image Regression in Functional Mixed Models

Hongxiao Zhu, SAMSI; Philip J. Brown, University of Kent, Canterbury; ✦Jeffrey S Morris, The University of Texas MD Anderson Cancer Center, , *jefmorris@mdanderson.org*

Key Words: Functional Data Analysis, Functional Mixed Models, Robust Methods, Outlier Detection, Image Analysis, Bayesian Methods

New methods have been developed in recent years to analyze functional and image data, many of which involve extensions of linear regression such as functional regression and functional mixed models.

Existing methods, however, tend to be sensitive to outliers, as no analogs to robust linear regression have been developed for the functional setting. Here, we discuss a unified Bayesian method for robust functional regression, whereby a functional response of unspecified form is regressed on a set of linear predictors. The method is developed within the general functional mixed model framework, which can simultaneously model multiple factors and accommodate between-function correlation induced by the experimental design. We demonstrate outstanding robustness properties, doing an excellent job estimating functional regression coefficients even in the presence of Cauchy errors and random effects, and yet not trading off much efficiency when the true likelihood is Gaussian. We also observed remarkable adaptive smoothing properties in our estimates of the fixed and random effect functions, which arise from an interaction of the robust likelihood and adaptive sparsity priors.

387 A Universe of Challenges: Development, Application, and Testing of Statistical Methods in Astronomy and Beyond ■●

Section on Physical and Engineering Sciences, Section on Nonparametric Statistics

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Introduction To Astrostatistics

✦Eric Feigelson, Penn State University, 407 Davey Lab, University Park, PA 16802 United States, *edf@astro.psu.edu*

Key Words: astrostatistics, image, Bayesian, truncation, time series, spatial processes

After a very brief review of the historical interplay between statistics and astronomy, this talk will discuss the state of astrostatistics today. Astronomers confront a bewildering variety of methodological challenges in analyzing and interpreting their data. Data can be in images, spectra, time series, and multivariate datasets. The samples can have either homogeneous or heterogeneous populations. Some datasets are sparse, while others reach petabyte scales. Astronomers need diverse methodologies including nonparametrics, regression, multivariate analysis and classification, time series analysis, spatial point processes, and image processing. Some particular challenges are common in astronomy: heteroscedastic measurement errors, truncation and censoring, and regression using complex models from astrophysical theory. Bayesian approaches to modeling are becoming popular. Astronomers are poorly trained in statistics, but interactions are improving with astrostatistical conferences, research groups, international organizations, and summer schools. Astrostatistics is thus emerging as an important cross-disciplinary enterprise engaging researchers from both communities.

Bayesian Planet Detection and Orbit Determination

✦Eric B. Ford, University of Florida, 211 Bryant Space Science Center, Gainesville, FL 32605, *ericbford@gmail.com*; Benjamin E. Nelson, University of Florida; Matthew J. Payne, University of Florida

Key Words: Astrostatistics, Population MCMC, GPU, Bayesian Parameter Estimation, Extrasolar Planet, Bayesian Model Selection

Astronomers detect planets orbiting distant stars based on observing the reflex motion of their host star. In the most common method for detecting planets, the basic observation data is a heteroscedastic, irregularly-sampled time series (the velocity of the host star towards or away from Earth). There is a simple physical model (i.e., star orbited by N planets), but the models are high-dimensional (~7*N) and highly non-linear (for N>1). Astronomers characterize planet masses and orbits via Markov chain Monte Carlo (MCMC) for Bayesian parameter estimation and model evaluation. For systems where the mutual planetary interactions are significant, the computation required for each model evaluation is significant (i.e., integrating a set of ~6*N ODEs describing the paths of each body). Additional physical constraints (e.g., long-term dynamical stability) can be imposed, but are orders of magnitude more computationally demanding. I will describe recent progress in the the development of population MCMC techniques to accelerate convergence and enable a high degree of parallelization (e.g., using Graphics Processing Units). Finally, I discuss future directions for astrostatistics.

Statistical Inference for Astronomical Populations from Truncated Data Sets

✦Brandon Kelly, Harvard-Smithsonian Center for Astrophysics, 60 Garden St., MS-06, Cambridge, MA 02138 US, *bckelly@cfa.harvard.edu*

Key Words: Bayesian Statistics, Astrophysics, Measurement Error, Truncation

Understanding demographics of astronomical populations and their evolution is often one of the primary goals of large astronomical surveys. However, this is not always a straightforward task in that the physical quantities, such as mass, that one is interested in are not directly measurable, but rather they are derived from measurable quantities, such as brightness, with uncertainty. Moreover, the situation is complicated by data truncation caused by brightness limits of telescopes. This complicates statistical inference on these populations, making it difficult to recover their evolution. In this talk I will discuss a Bayesian approach to this problem, as well as recent application of this approach to astronomical surveys. I will conclude by outlining areas for further improvement.

Facing Heteroscedastic Measurement Error in Astronomical Surveys

✦Chad Schafer, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 United States, *cschafer@stat.cmu.edu*

Key Words: astronomy, measurement error, density estimation, nonparametrics, truncation, likelihood-free inference

Current and forthcoming sky surveys force astronomers to face a difficult challenge: The estimation of the redshifts (a key proxy for distance in space-time) of astronomical objects using only limited, noisy observations of these objects. The resulting heteroscedastic measurement error strongly affects downstream inference procedures that require these quantities. We will focus on the implications for nonparametric density estimation in astronomical problems; this is an application that

already posed difficulties due to irregular truncation of the space of observables. We will discuss the potential of likelihood-free methods for handling such challenges.

388 Biometrics Showcase Session

WNAR, ENAR

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Learning Oncogenic Pathways from Binary Genomic Instability Data

✦Li Hsu, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., M2-B500, Seattle, WA 98109 USA, *lih@fhcrc.org*

Key Words: Conditional Dependence, Graphical Model, Lasso, Loss-of-Heterozygosity, Regularized logistic regression

Genomic instability, the propensity of aberrations in chromosomes, plays a critical role in the development of many diseases. High throughput genotyping experiments have been performed to study genomic instability in diseases. The output of such experiments can be summarized as high dimensional binary vectors, where each binary variable records aberration status at one marker locus. It is of keen interest to understand how aberrations may interact with each other, as it provides insight into the process of the disease development. In this talk, I will describe a novel method, LogitNet, for inferring such interactions among these aberration events. The method is based on penalized logistic regression with an extension to account for spatial correlation in the genomic instability data. I will present some simulation results to show that the proposed method performs well in the situations considered. Finally, I will illustrate the method using genomic instability data from breast cancer samples.

Regret Regression for Optimal Dynamic Treatment Regimes

✦Robin Henderson, Newcastle University, International United Kingdom, *Robin.Henderson@ncl.ac.uk*

Key Words: Anticoagulation, Causal inference, Path analysis, Diagnostics, Dynamic treatment

We consider optimal dynamic treatment regime determination in practice. Model building, checking and comparison have had little or no attention so far in this literature. The work is motivated by an application on long term treatment with anticoagulants, where optimal doses vary not just between people but also over time within subjects in response to short or long term changes in lifestyle. We propose a modelling and estimation strategy which incorporates the regret functions of Murphy (2003) into a regression model for observed responses, in a similar but more general manner to that proposed by Almirall et al (2009). Estimation is quick and diagnostics are available, meaning a variety of candidate models can be compared. The method is illustrated using simulation and the anticoagulation application, with emphasis on diagnostics, including wild bootstrap goodness-of-fit tests which do not require identically distributed observations.

Response-Adaptive Regression for Longitudinal Data

✦Shuang Wu, University of Rochester, 601 Elmwood Avenue, Rochester, NY 14642, *shuang_wu@urmc.rochester.edu*; Hans-Georg Mueller, University of California, Davis

Key Words: AIDS clinical trials, Functional linear regression, Growth curves, Prediction, Repeated measurements, Sparse data

We propose a response-adaptive model for functional linear regression, which is adapted to sparsely sampled longitudinal responses. Our method aims at predicting response trajectories and models the regression relationship by directly conditioning the sparse and irregular observations of the response on the predictor, which can be of scalar, vector or functional type. This obliterates the need to model the response trajectories, a task that is challenging for sparse longitudinal data and was previously required for functional regression implementations for longitudinal data. The proposed approach turns out to be superior compared to previous functional regression approaches in terms of prediction error. It encompasses a variety of regression settings that are relevant for the functional modeling of longitudinal data in the life sciences. The improved prediction of response trajectories with the proposed response-adaptive approach is illustrated for a longitudinal study of Kiwi weight growth and by an analysis of the dynamic relationship between viral load and CD4 cell counts observed in AIDS clinical trials.

389 Statistical Genetics ●

IMS, International Chinese Statistical Association, Section on Statistics in Epidemiology

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

The Many Facets of Genetic Data

✦Augustine Kong, deCODE Genetics, Sturlugata 8, IS-101, Reykjavik, Iceland, *augustine.kong@decode.is*

Key Words: genetic, information

With the rapid progress of technology, the genetic data being generated are becoming increasingly rich and complex. This creates a premium for novel analytical methods and sophisticated modelling that can extract the most information out of the data. At the same time, to have practical value, these methods have to be robust statistically and efficient computationally. In some instances, new statistical notions may be needed to appropriately capture the structure of the data and the complexity of the problems. Examples based on actual data and discoveries will be used for illustrations.

Association Mapping of Complex Traits in Dependent Samples

✦Mary Sara McPeek, University of Chicago Department of Statistics, Department of Human Genetics, , *msmcpeek@gmail.com*

Key Words: complex trait, population structure

One fundamental problem of interest is to identify genetic variants that contribute to observed variation in human complex traits. Large-scale genetic studies can have several sources of dependence, including longi-

tudinally-measured traits, population structure, relatedness among the sampled individuals and co-inheritance of nearby sites. The different types of dependence vary in the extent of information that is available on them, and this informs our approach to inference. Other characteristics of the data include missing information, and the need to analyze very large numbers of sites in a single study, which puts a premium on computational speed of the methods. We describe statistical methods for association mapping of human complex traits, in which we model and account for various types of dependence.

Gene-Set Analyses for Genome-Wide Association Studies (GWAS) Using Gene Ontology

✦Daniel J. Schaid, Mayo Clinic, Division of Biomedical Statistics and Informatics, Harwick 7, Rochester, 55905 USA, *schaid@mayo.edu*

Key Words: Gene Ontology, Gene Sets, Genome Wide Association, Score Statistics

Genome wide association studies (GWAS) measure hundreds of thousands of genetic markers (single nucleotide polymorphisms, SNPs) on large numbers of diseased cases and non-diseased controls, with most results reported according to the association of single SNPs with disease status. Although most studies find small odds ratios ranging 1.25 - 1.5 for SNPs, the benefit of GWAS can be enhanced by using prior information about how genes work together in biological pathways to create sets of genes. This presentation will discuss general strategies for scoring SNPs, combining these scores into gene-level scores, and then combining across genes in the same set. We use the publically available Gene Ontology to recursively create gene sets, capitalizing on its directed acyclic graph structure. Strengths and limitations of our approach will be discussed, as well as future research directions.

390 Computationally Intensive Modeling ■

Technometrics, Section on Statistical Computing

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Efficient MCMC Schemes for Computationally Expensive Posterior Distributions

✦David Nott, National University of Singapore, Department of Statistics and Applied Probability, 6 Science Drive 2, Singapore, International 117546 Singpaore, *standj@nus.edu.sg*

Key Words: Computer models, Hybrid Monte Carlo, Markov chain Monte Carlo, Parallel tempering

We consider Markov chain Monte Carlo (MCMC) computational schemes intended to minimize the number of evaluations of the posterior distribution in Bayesian inference when the posterior is computationally expensive to evaluate. Our motivation is Bayesian calibration of computationally expensive computer models. An algorithm suggested previously in the literature based on hybrid Monte Carlo and a Gaussian process approximation to the target distribution is extended in three ways. First, we consider combining the original method with tempering schemes. Second, we consider replacing the original target posterior distribution with the Gaussian process approximation, which requires less computation to evaluate. Third, we consider in the context of tempering schemes the replacement of the true target distribution with the approximation in the high temperature chains while retaining the true target in the lowest temperature chain. This retains the correct target distribution in the lowest temperature chain while avoiding the computational expense of running the computer model in moves involving the high temperatures.

Regression-Based Inverse Distance Weighting with Applications to Computer Experiments

✦Lulu Kang, Department of Applied Mathematics, Illinois Institute of Technology, Engineering 1 Building, Rm 208, 10 West 32nd Street, Chicago, IL 60616, *lkang2@math.iit.edu*; Roshan Joseph Vengazhiyil, School of Industrial and Systems Engineering, Georgia Institute of Technology

Key Words: confidence interval, Kriging, multivariate interpolation

Inverse distance weighting (IDW) is a simple method for multivariate interpolation but has poor prediction accuracy. In this article we show that the prediction accu- racy of IDW can be substantially improved by integrating it with a linear regression model. This new predictor is quite flexible, computationally efficient, and works well in problems having high dimensions and/or large data sets. We also develop a heuristic method for constructing confidence intervals for prediction.

391 The new world of data on human beings: Challenges and Solutions to Promoting Research while Ensuring Confidentiality ■●

Committee on Privacy and Confidentiality, Committee on Professional Ethics, Section on Government Statistics, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

The New World of Data on Human Beings: Challenges and Solutions to Promoting Research While Ensuring Confidentiality

✦Myron Gutmann, National Science Foundation, , *mgutmann@nsf.gov*; ✦Stephen E Fienberg, Carnegie Mellon University, , *fienberg@stat.cmu.edu*; ✦Ian Foster, University of Chicago, , *foster@anl.gov*; ✦Emmanuel Saez, University of California, Berkeley, , *saez@econ.berkeley.edu*; ✦Peter Elias, University of Warwick, , *peter.elias@warwick.ac.uk*

Key Words: Privacy, Confidentiality, Ethics, Data Access, Computational Data, CyberEnabled Data

This session is cosponsored by the Committee on Ethical Practice Advances in cyberinfrastructure have created a virtual deluge of new types of data ranging from new data on human interactions through digital imaging, sensors, and analytical instrumentation to new ways of collecting biological and geospatial information from survey respondents and to combining data from different sources, such as surveys and administrative records. In addition, new computational capacity has

emerged that facilitates the analysis of the data in terms of modeling and simulation with an unprecedented breadth and depth and scale. At the same time, new instrumentation provides unprecedented opportunity for researchers to advance scientific understanding through collaboration with colleagues around the globe. Major funding agencies have recognized the importance of these new data for advancing social science research, as well as the importance of providing access to data for research - and have mandated data management and open access policies to promote the dissemination of ideas.

392 Recent Development of High-dimensional Statistical Learning ■●

Section on Statistical Learning and Data Mining, International Chinese Statistical Association, International Indian Statistical Association, Section on Statistical Computing

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Penalized Bregman Divergence Estimation Via Coordinate Descent

✦Chunming Zhang, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, *cmzhang@stat.wisc.edu*

Key Words: Bregman divergence, LARS algorithm, quasi-likelihood, sparsity, variable selection

Variable selection via penalized estimation is appealing for dimension reduction. For penalized linear regression, Efron, etal (2004) introduced the LARS algorithm. Recently, the coordinate descent (CD) algorithm was developed by Friedman, etal (2007) for penalized linear regression and penalized logistic regression and was shown to gain computational superiority. This paper explores the CD algorithm to penalized Bregman divergence (BD) estimation for a broader class of models, including not only the generalized linear model, which has been well studied in the literature on penalization, but also the quasi-likelihood model, which has been less developed. Simulation study and real data application illustrate the performances of the CD and LARS algorithms in regression estimation, variable selection and classification procedure when the number of explanatory variables is large in comparison to the sample size.

Non-Concave Penalized Likelihood With Np-Dimensionality

Jianqing Fan, Princeton University; ✦Jinchi Lv, University of Southern California, Information and Operations Management Department, Marshall School of Business, Los Angeles, CA 90089 USA, *jinchilv@marshall.usc.edu*

Key Words: High-dimensional variable selection, Non-concave penalized likelihood, Folded-concave penalty, Oracle property, Weak oracle property, Lasso; SCAD

Penalized likelihood methods are fundamental to ultra-high dimensional variable selection. How high dimensionality such methods can handle remains largely unknown. In this paper, we show that in the context of generalized linear models, such methods possess model selection consistency with oracle properties even for dimensionality of Non-Polynomial (NP) order of sample size, for a class of penalized likelihood approaches using folded-concave penalty functions, which were introduced to ameliorate the bias problems of convex penalty functions. This fills a long-standing gap in the literature where the dimensionality is allowed to grow slowly with the sample size. Our results are also applicable to penalized likelihood with the L_1-penalty, which is a convex function at the boundary of the class of folded-concave penalty functions under consideration. The coordinate optimization is implemented for finding the solution paths, whose performance is evaluated by a few simulation examples and the real data analysis.

Group Iterative Sure Independent Screening

✦Ning Hao, The University of Arizona, 617 N. Santa Rita Ave., Tucson, AZ 85721, *nhao@math.arizona.edu*

Key Words: Genome Wide Association Study, Group Selection, Group ISIS, High Dimensionality

We consider regression problems in which the predictors possess a group structure. The goal is to select important groups efficiently in the ultra-high dimensional setting, when the number of groups is much larger than the sample size. The ultrahigh setting is common in contemporary data sets. In Genome-Wide Association Study, important SNPs and genes are to be selected in a pool consisting of tens of thousands of SNPs within thousands of genes while the sample size is usually about a few hundreds. Penalization approaches such as group LASSO and group MCP are not efficient in solving ultrahigh dimensional problems. We use correlation learning approach and generalize the Iterative Sure Independent Screening (ISIS) procedure to the setting of group selection. Group ISIS is proposed for group selection. An application in Genome-Wide Association Study is shown as well.

Scaled Sparse Linear Regression

✦Tingni Sun, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, *tingni@stat.rutgers.edu*; Cun-Hui Zhang, Rutgers University

Key Words: variance estimation, convex minimization, asymptotic normality, linear regression

Scaled sparse linear regression jointly estimates the regression coefficients and noise level in a linear model. It chooses an equilibrium with a sparse regression method by iteratively estimating the noise level via the mean residual squares and scaling the penalty in proportion to the estimated noise level. The iterative algorithm costs nearly nothing beyond the computation of a path of the sparse regression estimator for penalty levels above a threshold. For the scaled Lasso, the algorithm is a gradient descent in a convex minimization of a penalized joint loss function for the regression coefficients and noise level. Under mild regularity conditions, we prove that the method yields simultaneously an estimator for the noise level and an estimated coefficient vector in the Lasso path satisfying certain oracle inequalities. These oracle inequalities provide sufficient conditions for the consistency and asymptotic normality of the estimator for the noise level, including cases where the number of variables is of greater order than the sample size. Numerical results demonstrate the superior performance of the proposed method over an earlier proposal of joint convex minimization.

Neyman-Pearson Paradigm in Binary Classification

✦Xin Tong, Princeton University, , *xtong@princeton.edu*

Key Words: binary classification, Neyman-Pearson paradigm, anomaly detection, empirical constraint, empirical risk minimization

Motivated by problems of anomaly detection, this paper implements the Neyman-Pearson paradigm to deal with asymmetric errors in binary classification with a convex loss. Given a finite collection of classifiers, we combine them and obtain a new classifier $\hat{f}$ that satisfies simultaneously the two following properties with high probability: (\rmnum{1}), type \Rmnum{1} error of $\hat{f}$ is below a pre-specified level α; (\rmnum{2}), $\hat{f}$ has type \Rmnum{2} error close to minimum under the type \Rmnum{1} constraint. The classifier $\hat{f}$ is obtained by solving an optimization problem with an empirical objective and an empirical constraint. Moreover, we address the case where we have more observations in one class than the other, as it is in anomaly detection problems.

393 Linking Administrative Data to Survey Data: Implications for Consent ●

Section on Government Statistics, Section on Health Policy Statistics, Section on Risk Analysis, Section on Survey Research Methods, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Jumping The Informed Consent Hurdle: Federal Agency Experiences

✦Mary Frazier, Census Bureau, 4700 Silver Hll Rd, Washington, DC 20233, *mary.b.frazier@census.gov*; Cynthia Nickerson, USDA Economic Research Service

Key Words: administrative records, linking, informed consent

The federal statistical community has understood the benefits of linking survey data and administrative records for many years. Likewise, federal agencies have been sharing and linking survey and administrative data for many years. Yet, agencies often cite difficulties with obtaining access to administrative data or permission to link data for statistical purposes. One barrier, critical for statstical agencies that rely on the public to voluntarily provide data, is informed consent. Current law, regulation and guidance include fairly clear requirements regarding what individuals should be told about the purposes and use of data being collected. However, there is considerably less guidance on what individuals must be told when data are shared with other agencies and/or linked with other data sources for statistical uses. Members from the FSCM Adminstrative Records and the FCSM Privacy Subcommittees formed a working group to examine this issue and how it impacts agencies ability to obtain and use administrative records for statistical use. This discussion reviews the results of their effort including examples of notice with a focus on how the recommendations may ease the challenges.

Historical Linkage Of Tax Data On Labour: A Case Applied To The Living In Canada Survey Pilot

✦Manon Langevin, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6 Canada, *Manon.Langevin@statcan.gc.ca*; Andrew Heisz, Statistics Canada

Key Words: historical linkage, retrospective linkage, tax data, administrative data, matched data, linkage rate

Matching data is a common practice that allows reducing response burden, as well as improving the quality of the information collected from the respondents when the linkage method does not introduce a bias. However, historical linkage, which consists in linking external records from past years with the initial wave of a survey, is a method relatively unknown and that was, up to now, never used in a social survey at Statistics Canada. The present paper describes the method used for linking the records from the Living in Canada survey pilot with retrospective tax data on labour and income (T4 and T1 forms). We will also discuss the characteristics of the records where linkage was a success or not, and the impact of these characteristics on the data quality and the linkage rates through time. To demonstrate the new possibilities of analysis brought by historical data matching, the study also compares the profile of earnings according the age and sex of different cohorts based on the year of birth.

Measuring Non-Consent Biases In The German Labour Market And Social Security Study

Joseph Sakshaug, University of Michigan; Frauke Kreuter, University of Maryland; ✦Stefan Bender,

Key Words: data linkage, consent bias, measurement error, nonresponse bias

Administrative data are increasingly being linked to survey data to enhance the survey data and increase research opportunities for data users. A necessary prerequisite to performing direct record linkage is obtaining informed consent from respondents. Respondent willingness to consent is not universal. Several studies have found systematic differences between consenters and non-consenters on survey variables, including socio-demographic characteristics, raising concerns that inferences obtained from linked data sources may be biased. We estimate non-consent biases in the German Labour Market and Social Security Study (PASS). About 80% of PASS respondents provided informed consent to link survey, employment, and benefit recipiency records. With permission from the German Institute for Employment Research, we analyzed administrative records for both consenters and non-consenters and estimated non-consent biases for several administrative variables. In addition, we estimate nonresponse and measurement error biases for the same variables. This paper concludes with a comparison of the different error sources and their relative contributions to the overall error in the linked data.

Record Linkage In The Survey Of Health Ageing And Retirement In Europe

Annelies Blom, SHARE, MEA, Mannheim University; ✦Ulrich Krieger, MEA, University of Mannheim, L13, 17,, Mannheim, Germany, *krieger@mea.uni-mannheim.de*; Julie M. Korbmacher, SHARE, MEA, Mannheim University

Key Words: record linkage, administrative data, consent, interviewer effects

Linking survey data with data from administrative records is of interest for both substantive researchers (because it enriches the data) and survey methodologists (because it enables research into measurement error and selectivity of the linked sample). The German sample of the Survey of Health, Ageing and Retirement in Europe (SHARE) asks respondents for permission to link their survey data to individual records from the German Pension Fund. In wave 3 (2008) all longitudinal sample members were asked for consent to linkage via their social security number. In wave 4 (2010) an additional three-quarters of the refresher sample was asked for consent to linkage. In addition, an interviewer questionnaire was administered collect data that might explain interviewer effects on consent propensity. Our analyses look into differences in the selectivity of the linked samples across the two waves and the role of interviewers thereupon.

394 Bernie Harris: A Life in Statistics

Section on Risk Analysis, International Indian Statistical Association, Section on Statistics in Defense and National Security

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Bernie Harris And Systems Reliability

✦Nozer Singpurwalla, The George Washington University, Department of Statistics, 2140 Pennsylvania Avenue, Washington, DC 20052, *nozer@gwu.edu*

Key Words: Propensity, Survivability, Networks, Interdependence

Bernie Harris contributed much to various topics in mathematics, statistics, and probability. One of the topics in which he made a signal contribution pertained to interval estimation of system reliability. He used some powerful technical machinery to obtain his results. Bernie's approach was frequentist. In this talk I take a Bayesian approach to the problem and make the argument that the reliability of the system is to be viewed as a propensity and that what one needs to be concerned about is the uncertainty about this propensity as measured by a probability. I also make the case that independence in system survivability assessments should be argued hierarchically.

Some Memories Of Bernie Harris

✦Arthur Fries, Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311, *afries@ida.org*

Key Words: Decision Theory, Defense, Counter-Terrorism

This talk features some personal reminiscences of the author's interactions with Bernie Harris over the last 30 years, beginning with Decision Theory courses that Bernie taught at the University of Wisconsin and extending through various encounters at professional forums over the years.

Bernard Harris' Contributions To Engineering Statistics: Reliability, Tolerance Limits, And Service To The U.S. Army

Bernard (Bernie) Harris was prominent in both theoretical and applied statistics. But those of us fortunate enough to have known Bernie will never forget his kindness and his wonderful (and probably unique) sense of humor. This presentation will review his contributions to engineering statistics, and his service to the U.S. Army at the Mathematics Research Center and as a consultant to the Army Materials Technology Laboratory. Along the way, I hope to illustrate through my own experiences the extra-statistical aspects of Bernie's personality, which made him very special and which will be greatly missed.

Bernie Harris' Contributions To Cluster Analysis

✦Stanley L Sclove, Information & Decision Sciences Dept., Univ. of Illinois at Chicago, 1080 HILLCREST RD, GLENCOE, IL 60022-1216, *slsclove@uic.edu*

Key Words: cluster analysis, clustering, graph theory, combinatorics

This talk will describe some of Bernie Harris' contributions to cluster analysis. The problem of cluster analysis is, given observation vectors on a sample of objects or individuals, group them. Harris focused on the detection of clustering by viewing the situation via the number of edges and cliques in graphs, in particular, via the expected number by chance alone. The talk also includes personal reminiscences, collected over years of interaction at various conferences, esp. the Classification Society (Bernie was an active member for a number of years and organized and hosted one of our meetings) and sessions of the Risk Analysis Section at JSM (Bernie was a founding member).

395 Advanced Statistical Method for Marketing Research ■

Section on Statistics and Marketing, Section for Statistical Programmers and Analysts

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Demography, Geography And Marketing: Telmar CentabÆ-The Largest Us Census Based Study From Telmar. Media And Marketing Studies Can Now Be Cross-Tabbed At Custom Geographic Levels

✦Igor Mandel, Telmar Group Inc., 470 Park avenue South, New York, NY 10016, *igor.mandel@gmail.com*

Key Words: advertising, media planning, demographic database, Census, Data Fusion and Ascription, Telmar

Telmar Group Inc. has created and made available for the first time a single respondent database combining the most reliable source of demographic information - annual Public-Use Microdata Samples (PUMS) data in the American Community Survey by the US Census Bureau and combined it with updated county data from independent census data marketers, to produce over 3 million of real respondents (without their addresses), distributed across 3,000 counties and soon across 40,000 zip codes. It provides marketers with the unique opportunity to have an annually updated, highly reliable, geographically specified crosstab-able data base with 150 different variables, covering all traditional metrics used in advertising media planning. It will also

serve as the most reliable hub for all forms of data integration, such as Fusion, Multi-Basing, Weighted Profile Analysis and Ascriptions. These methods, together with Sample Balancing procedures used, will be discussed. Presentation talks also about how this database was created and demonstrates some findings from it including correlations between variables; some of them are rather surprising and dismiss or correct popular beliefs.

Implications Of Using Survey Routers In Marketing Research

✦Shon Magnan, GfK Custom Research North America, 8401 Golden Valley Road, P.O. Box 27900, Minneapolis, MN 55427, *shon.magnan@gfk.com*

Key Words: marketing research, sampling, routers

Survey routers in online research are used to assign incoming respondents to open surveys for which they are likely to quality. Router benefits touted include enhanced efficiency, support for the fielding of surveys to low-incidence populations and respondent experience by minimizing the number of screen-outs or "quota full" messages. Routers however also can introduce sampling bias especially if priority routers are used instead of random routers. Priority routers assign respondents who quality for multiple surveys to one survey in preference over all others. Prioritizing surveys may be efficient but can also create selection bias as the set of respondents sent to unprioritized surveys differ from the potential respondents who enter the router. Recent research suggests that although there is indeed bias introduced, the bias does not affect the outcome of the research. This presentation discusses the issue of survey routers, the potential bias introduced and explores the affect of this bias via a series of simulations.

Simple Pricing Models With Monotonicity Constraints In Marketing Research

✦Joseph Retzer, MarketTools Inc., 2019 E. River Rd., Grafton, WI 53024, *retzerjj@gmail.com*; Patrick St. John , MarketTools Inc.

Key Words: Hierarchical Bayes, Pricing models, Van Westendorp, Gabor Granger, Monotonicity Constraints, Discrete Choice

Managerial decisions utilizing pricing research results are ubiquitous in marketing mix analysis. This presentation begins with a review of standard approaches designed to determine general price perceptions. The models discussed will include the Van Westendorp and Gabor Granger pricing models. While both models have certain strengths, they also suffer from the lack of elicitation of consumer intention in a competitive context. Subsequent analysis will focus on discrete choice pricing experiments via aggregate estimation both with and without monotonicity constraints. While this approach introduces a competitive context, it ignores respondent heterogeneity, which in turn limits predictive performance. Finally, we examine results generated by discrete choice experiments where price level utilities are approximated using hierarchical Bayesian estimation. The estimates will be generated with and without various monotonicity constraints. Data will be collected for all models using blocks of randomly assigned consumers surveyed on the same product(s). This approach will facilitate a general comparison and critique of each method.

Causal Inferences with Linear Matching

✦Volker Bosch, GfK SE, Nordwestring 101, Nuremberg, International 90419 Germany, *volker.bosch@gfk.com*

Key Words: Matching, Bias Reduction, Propensity Score Matching

In a controlled study design experimental and control group must be structurally identical to allow for causal inferences. However, in real life problems structures differ greatly due to selection bias. Thus, the groups must be statistically matched. In academia, state of the art is Propensity Score Matching where bias due to structural variables including interactions thereof is reduced to the same imperfect extent. In contrast, in market research matching restricted to the main effect level is preferred assuming that interaction bias is negligible. In any case, high test power for group comparison should be preserved. We developed a linear matching algorithm that perfectly aligns the structures of the test and the control group but ignores interactions of structural variables unless explicitly specified. Importantly, the algorithm is optimal under the given constraints with respect to statistical power. Studies on customer cards utilizing household panel data and on promotion effectiveness utilizing sales data demonstrate the advantages of the approach. Moreover, the results indicate that bias due to unspecified interactions is indeed negligible.

Dual PLS Analysis

✦Stan Lipovetsky, GfK Custom Research North America, 8401 golden Valley Rd., Minneapolis, 55427, *stan.lipovetsky@gfk.com*

Key Words: Dual Multivariate Statistical Analysis, Partial Least Squares, Inter-Battery Factor Analysis, Robust Canonical Correlations

A new multivariate statistical technique is obtained for comparing and combining two or more data sets each of which has a different number of respondents but the same variables. This approach can be considered as dual to such techniques as partial least squares, also known as inter-battery factor analysis and robust canonical correlation analysis for two data sets. It is shown that the problem can be reduced to the eigenproblem of the product of correlation matrices of each data set. The technique is generalized to three or more data sets in an eigenproblem of block-matrices of the correlations within each data set. This type of multivariate analysis can serve various practical problems of integration of data obtained from heterogeneous sources, particularly, for data merging in constructing data warehouses.

396 Recent Developments in Statistical Ecology ■●

Section on Statistics and the Environment, WNAR, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Model-Based Clustering for Metapopulation Trends

✦Devin Johnson, NOAA National Marine Mammal Laboratory, 7600 Sand Point Way NE, 7600 Sand Point Way NE, Seattle, WA 98115 United States of America, *devin.johnsonn@noaa.gov*

Key Words: Metapopulation, trends, Dirichlet process, Markov random fields, abundance, reversible jump MCMC

We consider a model-based clustering approach to examining abundance trends in a metapopulation. When examining trends for an animal population with management goals in mind one is often interested in those segments of the population that behave similarly to one another with respect to abundance. Our approach to trend analysis incorporates a clustering approach that is an extension of the classic Chinese Restaurant Process, and the associated Dirichlet process prior, which allows for inclusion of similarity covariates between sites. Both overall linear trend an smooth variations around the overall trend are considered in this analysis. To demonstrate the approach we examine long term trends in northern fur seal pup production at nineteen rookeries in the Pribilof Islands, Alaska.

Hierarchical Modeling Approaches for Modeling Ecological Communities

✦Ali Arab, Georgetown University, Department of Mathematics and Statistics, 322 St. Mary's, 37th and O streets, Washington, DC 20057, *aa577@georgetown.edu*

Key Words: hierarchical models, ecological communities, count data, multivariate distributions

The problem of modeling dependence structures and complex relationships among species in an ecological community is a challenging task and requires careful implementation of statistical methods that are both realistic and plausible (i.e., easy to implement and interpret). Most conventional methods used for modeling community data have serious shortcomings (e.g., unrealistic or restrictive assumptions on data and model). For example, assumptions of normality and independence often do not reflect the nature of the data (often count data) and the relationships among species (often better explained using multivariate distributions). Hierarchical modeling approaches discussed in this work provide a more flexible and plausible but also easy to implement alternative for modeling ecological community data. The hierarchical modeling framework seems natural for modeling complex relationships among species (e.g., predator-prey-competitor structures). The modeling approaches discussed will be illustrated using data from ecological communities (e.g., food web data, fish communities).

Estimating Abundance-Based Patterns of Species Co-Occurrence Using Phylogenetic Data and Spatial Covariates

✦Robert M. Dorazio, Southeast Ecological Science Center, U.S. Geological Survey, Department of Statistics, University of Florida, Gainesville, FL 32611-0339, *bdorazio@usgs.gov*; Edward F. Connor, San Francisco State University

Key Words: abundance, co-occurrence, detectability, foraging guilds, point counts

We develop a statistical model to estimate the abundances of species encountered while surveying a set of ecologically relevant locations -- as in a metacommunity of species. In the model we assume that abundances of related species (e.g., species of the same foraging guild) are correlated. We also assume that abundances vary among locations owing to systematic and stochastic sources of heterogeneity. For example, if abundances differ among locations owing to differences in habitat, then measures of habitat can be included in the model as covariates. Naturally, the quantitative effects of these covariates are assumed vary among species. In the model we also account for the effects of detectability on the observed counts of each species. This aspect of the model is especially important for rare species that may be difficult to detect in multi-species surveys. We illustrate the model using point counts of avian species obtained while sampling a community of forest birds during the breeding season.

Predicting Infectious Disease Outbreak Carried by Migratory Waterfowl

✦Jacob J. Oleson, The University of Iowa, Department of Biostatistics, 200 Hawkins Drive, C22 GH, Iowa City, IA 52242-1009, *jacob-oleson@uiowa.edu*; Christopher K. Wikle, University of Missouri

Key Words: Bayesian, Functional, Hierarchical, Markov chain Monte Carlo, Risk, Spatio-temporal

Given the uncertainties associated with vector-born infectious diseases, it is critical to develop statistical models to address how and when an infectious disease could spread throughout a region such as the United States. Modeling spatio-temporal data of this type is inherently difficult given the uncertainty associated with observations, complexity of the dynamics, high dimensionality of the underlying process, and the presence of excessive zeros. The spatio-temporal dynamics of a waterfowl migration are developed by way of a novel two-tiered functional temporal and spatial dimension reduction procedure that captures spatial and seasonal trends, as well as regional dynamics. Furthermore, the model relates the migration to a population of poultry farms that are known to be susceptible to such diseases, and is one of the possible avenues towards transmission to domestic poultry and humans. The result is a predictive distribution of those counties containing poultry farms that are at the greatest risk of having the infectious disease infiltrate their flocks assuming that the migratory population was infected. The model fits into the hierarchical Bayesian framework.

Velocity-Based Movement Modeling for Individual and Population-Level Inference

✦Ephraim M Hanks, Colorado State University, CO , *hanks@stat.colostate.edu*; Mevin B Hooten, USGS Colorado Cooperative Fish and Wildlife Research Unit

Key Words: Animal Movement, Reversible Jump MCMC, Change Point Model

Understanding environmental drivers of animal movement and resource selection provides important information about the ecology of the animal, but an animal's response to the environment is not typically constant in time. We present a velocity-based approach for modeling animal movement that allows for temporal heterogeneity in an animal's response to the environment, allows for temporal irregularity in telemetry data, accounts for the uncertainty in the location information, and scales up naturally to population-level inference. We illustrate this approach through a study of northern fur seal (Callorhinus ursinus) movement in the Bering Sea, AK.

397 New Developments in American Community Survey Methods and Mapping Applications

Section on Survey Research Methods, Section on Government Statistics, Section on Health Policy Statistics, Social Statistics Section, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Using Imputation Methods To Improve The American Community Survey Estimates Of The Group Quarters Population For Small Geographies

Mark Asiala, U.S. Census Bureau; ✦Michael Beaghen, U.S. Census Bureau, 6641 Wakefield Drive, #518, Alexandria, VA 22307, *michael.a.beaghen@census.gov*; Alfredo Navarro, U.S. Census Bureau

Key Words: sample design, small area estimation

This paper describes the background and methodology of a Census Bureau program under development to improve the American Community Survey (ACS) estimates of the group quarters (GQ) population for small areas. What motivates this work is that while the ACS GQ sample was designed to produce estimates at the state-level, the estimates of the GQ population contribute to ACS estimates of the total resident population for substate areas such as counties and tracts. Consequently, there are small geographies which either do not have GQ sample or have GQ sample that is not representative of the area, which can lead to distorted estimates of characteristics and/or total population for these geographies. The approach taken is to impute whole person records (and weight them appropriately) to GQ facilities which appear on the sampling frame but were not selected into sample.

Using Statistical Process Control Techniques In The American Community Survey

✦Steven Hefter, U.S. Census Bureau, Washington, DC 20233, *steven.p.hefter@census.gov*; Erica Marquette, U.S. Census Bureau

Key Words: ACS, Statistical Process Control, Error

Each year the American Community Survey collects data about roughly two million housing units, 4.5 million people in the household population, and 150,000 people in group quarters facilities. We have begun developing automated statistical process control methods to uncover potential errors in the data. Several methodologies are being used to investigate responses from all three data collection modes (mail, Computer Assisted Telephone Interview (CATI), and Computer Assisted Personal Interview (CAPI)) using traditional Shewhart charts. For mail, we compare the individual responses of each question at various levels of geography. For CATI, we use similar methods, but also compare data for each telephone center. For the CAPI data collection mode we concentrated our initial efforts on Field Representative (FR) item missing data rates and compare each FR to all FRs within clusters of counties. Our paper presents the details of the methodology, and several examples and results, as well as discussion of the inherent challenges and obstacles faced when applying traditional process control methods to a large-scale, multi-mode, demographic survey.

Incorporating A Finite Population Correction Factor In American Community Survey Variance Estimates

✦Michael Starsinic, U.S. Census Bureau, 14607 London Lane, Bowie, MD 20715, *michael.d.starsinic@census.gov*

Key Words: American Community Survey, variance estimation, finite population correction

The American Community Survey (ACS) produced its first nationwide 5-year estimates in 2010, using sample data from 2005 through 2009. With five years' worth of sample, the combined sample size in some areas would be large enough that a finite population correction (FPC) factor might have a noticeable impact on variances. This paper will discuss the methodology used to incorporate an FPC factor into the 5-year ACS variance estimates, and how the method was adapted to account for the subsampling of nonrespondents . Results comparing the impact on the variance of using the FPC across a broad spectrum of estimates and geographic areas will also be presented. Preliminary work indicated improvements in the standard error estimates of between two and four percent could be achieved.

Improving Weighted Person-Level Estimates From The American Community Survey'S Public Use Microdata Sample

✦Bryan Dale Garrett, U.S. Census Bureau, 2713 Curry Drive, Adelphi, MD 20783-1725, *B.Dale.Garrett@census.gov*

Key Words: American Community Survey, weighting, public use microdata

The Public Use Microdata Sample (PUMS) files contain records for a subsample of the housing units and persons of the American Community Survey (ACS) annual sample. A weighting process was introduced for the 2009 PUMS that expanded the raking matrix to include more demographic controls and family equalization with the goal of forcing more consistency between the PUMS and the ACS full sample estimates. This paper discusses the preliminary research, the trade-offs of doing the weighting at the state versus PUMA levels, and some of the impact on estimates of the new weighting procedure.

Mapping American Community Survey Data and Beyond

✦Nancy K. Torrieri, U.S. Census Bureau, American Community Survey Office, Suitland, MD 20746, *nancy.k.torrieri@census.gov*; Michael Ratcliffe, U.S. Census Bureau; David Wong, George Mason University

Key Words: Census Bureau, American Community Survey, Mapping

Maps are a frequently used tool to portray the Census Bureau's data and highlight spatial patterns that provide context and significance for the characteristics displayed. Casual users of maps of statistical data may not look past what is interesting visually to analyze the underlying data that a map depicts. However, that does not absolve the mapmaker of the responsibility for informing users of the statistical qualities associated with the mapped values. The Census Bureau set new standards for communicating the statistical qualities of estimates from the American Community Survey (ACS) by including information on margins of er-

ror associated with every ACS estimate. It is appropriate that the Census Bureau also promote the inclusion of this information in its map products. The Census Bureau is seeking input from the spatial science community to meet this goal. This paper describes how geographic information system tools used to map ACS data could be enhanced to incorporate information on margins of error. It also provides an update on cartographic activities and research at the Census Bureau and describes how they may contribute to improving the Census Bureau's maps in the future.

398 Analysis of Test-Retest Reliability/ Agreement Studies ■

Biopharmaceutical Section, Section on Statistics in Epidemiology
Tuesday, August 2, 2:00 p.m.–3:50 p.m.

An Overview Of Different Methods To Assess Agreement In Early Phase Clinical Trials

✦Radha Railkar, Merck, UG1CD-44, 351 N. Sumneytown Pike, North Wales, PA 19454, *radha_railkar@merck.com*; Richard Baumgartner, Merck & Co., Inc.; Dai Feng, Merck & Co., Inc.; Cynthia Gargano, Merck; Patrick Larson, Merck; Lori Mixson, Merck

Key Words: agreement, biomarker, intraclass correlation coefficient, concordance correlation coefficient, within subject coefficient of variation

Early phase clinical trials are often undertaken to evaluate new measurement techniques or biomarkers that could help to rapidly identify clinical target activity of new compounds with smaller numbers of patients/subjects. In these studies it is important to establish that the new measurement technique or biomarker agrees with an existing gold standard or that 2 or more repeat measurements obtained using the new technique or biomarker, agree with one another (i.e., the technique or biomarker is repeatable). Common methods to assess agreement include the intraclass correlation coefficient (ICC), the concordance correlation coefficient (CCC), the within subject coefficient of variation, and graphical methods such as the Bland-Altman plot. We discuss the pros and cons of the various methods. Additionally, multiple models and methods have been proposed to estimate the ICC and CCC. We propose that estimating the ICC/CCC from a 2-way mixed effects model with the repeat measurements as fixed effects is appropriate for evaluating agreement in small studies. Simulation studies were performed to compare the different estimation methods for ICC/CCC and our recommendations are presented.

On The Bayesian Credible Intervals For Intraclass Correlation Coefficients With Small Number Of Raters

✦Dai Feng, Merck & Co., Inc., , *dai_feng@merck.com*; Valdimir Svetnik, Merck & Co., Inc.; Alexandre Coimbra, Merck & Co., Inc.; Richard Baumgartner, Merck & Co., Inc.

Key Words: Intralclass correlation coefficient, small number of raters, Bayesian credible interval

In drug development, an important task is to study reproducibility or test-retest reliability. Intraclass correlation coefficient (ICC) is a metric that has been widely used to assess the reproducibility. In particular, the ICC obtained from a mixed effects model with fixed rater effects is recommended in the situation with a small number of raters, which is a typical setup in early drug development studies. To calculate confidence intervals (CIs) for the ICC, various frequentist methods have been proposed. They include methods based on second and higher moment approximations, the delta method, and others. We propose using a Bayesian method with a Jeffreys' prior to obtain the credible sets. When there are two raters, the independent samples can be generated from constructive posteriors and obtained very quickly using vectorized computation in R. Judging by simulation studies and results on real EEG datasets, the Bayesian approach is at least comparable with and sometimes better than frequentist approaches based on different frequentist properties. The Bayesian method should be considered as an alternative for ICC CI calculation in early drug development.

Concordance Correlation Coefficient Decomposed into the Product of Precision and Accuracy

✦Christopher Tong, USDA Center for Veterinary Biologics, 1920 Dayton Avenue, Ames, IA 50010, *christopher.h.tong@aphis.usda.gov*

Key Words: Concordance correlation, Pearson correlation, Agreement, Method comparison

The concordance correlation coefficient (CCC) was introduced by Lin (Biometrics, 45: 255-268, 1989) as an index of agreement for paired measurements. It may be written as the product of the Pearson (product-moment) correlation and a bias correction factor. Lin calls these factors measures of precision and accuracy, respectively. The Pearson correlation "measures how far each observation deviated from the best-fit line" while the bias correction factor "measures how far the best-fit line deviates from" the 45 degree line of agreement, in Lin's words. I demonstrate that these claims can be misleading. Loh (J. Educ. Stat., 12: 235-239, 1987) showed that the Pearson correlation is not simply a measure of clustering of data around the best-fit line. The Pearson correlation is in fact sensitive to the departure of the best fit line from the line of agreement. Thus the Pearson correlation fails to be a pure measure of precision on the raw data. Alternative interpretations are discussed.

Finding A Matching Cutoff

✦Charles Tan, Pfizer Inc., 1111 Hunt Seat Drive, Lower Gwynedd, PA 19002, *charles.y.tan@pfizer.com*

Key Words: Cutoff, Assay, Precision, Semi-parametric

Sometimes patient's disease status or clinical response is defined by whether certain laboratory test crosses a threshold. For example, a patient is considered protected by pneumococcal vaccine if his or her post immunization serum antibody levels are above a cutoff. When a new assay measuring the same analyte or related analyte within the same biological pathway is developed, there is a need to determine a cutoff that best matches the cutoff in the original assay in terms of comparable ability to classify patients. For example, a cutoff for opsonophagocytic assay (OPA) which measures functional antibody level is sought to match known cutoff for the total IgG assay which measures circulating antibody level via binding. This presentation will describe a semi-

parametric method that utilizes the precision information from both assays. Precision is an important assay performance metric that has been typically ignored in this context. The semi-parametric nature of the method also allows flexibility when the two assays measure related, rather than identical, analyte.

399 Recent Advances in Congnitive Assesment ■●

ENAR, Section on Health Policy Statistics, Section on Statistics in Epidemiology

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Theory of Self-Learning Q-Matrix

✦Jingchen Liu, Columbia University, 1255 Amsterdam Ave, Room 1030, New York, NY 10027 USA, *jcliu@stat.columbia.edu*; Gongjun Xu, Columbia University; Zhiliang Ying, Columbia University

Key Words: Cognitive assessment, diagnostic classification model, Q-matrix, self-learning, consistency

Cognitive assessment is a growing area in psychological and educational measurement, where tests are given to assess mastery/deficiency of attributes or skills. A key issue is the correct identification of attributes associated with items in a test. In this paper, we set up a mathematical framework under which theoretical properties may be discussed. We establish sufficient conditions to ensure that the attributes required by each item are learnable from the data.

Cognitive Diagnosis Models with Longitudinal Growth Curves for Skill Knowledge

✦Elizabeth Ayers, University of California, Berkeley, CA , *eayers@berkeley.edu*; Sophia Rabe-Hesketh, University of California, Berkeley

Key Words: Cognitive Diagnosis, Growth Curve Modeling, Longitudinal

In recent years, a number of cognitive diagnosis models have become a popular means of estimating student skill knowledge. However these models treat responses as though they are from a single time point. When data is collected throughout a school year, we expect student skill knowledge at different times to be dependent within students and the probability of skill mastery to increase over time as students learn. We have developed longitudinal cognitive growth curve models to account for the within-student dependence, as well as understand the variability in learning and how this depends on explanatory variables. The relationship between the latent binary skill knowledge indicators and the item responses is modeled as a DINA model. A logistic regression model is specified for the latent skill knowledge indicators with student characteristics and time as covariates and with a student-level random intercept and random slope of time. The model is estimated using Markov chain Monte Carlo in WinBUGS. Simulation studies show good parameter recovery. The model will be applied to data from the ASSISTment tutor, an online mathematics tutor used by eighth graders in Massachusetts.

Do Diagnostic Models Hold More Promise Than They Deliver?

✦Matthias von Davier, ETS, Rosedale Road, Princeton, NJ 08541 USA, *mvondavier@ets.org*

Key Words: Latent class analysis, Diagnostic classification models, Categorical data analysis, Discrete latent variables

This talk will discuss recently developed models for complex student response data utilizing multidimensional discrete latent variables. These variables represent latent mastery/non-mastery levels which have to be inferred from observed responses to a series of items. Models of this type have been developed as extensions of latent class analysis to multiple classifications, and are nowadays often referred to as models for cognitive diagnosis, or diagnostic classification models. The class of models presented here is suitable for binary (correct/incorrect) and ordinal (partial credit) responses and allows modeling the latent structure as a combination of quantitative and qualitative latent variables. Applications of this general diagnostic modeling framework to analysis of test and questionnaire data from areas such as cognitive skills assessments, large scale surveys, longitudinal analysis, and the analysis of multidimensional personality tests (Big 5) will be discussed. The presentation will give an overview of the field, and will talk about promised advantages of using these models as well as limitations of these models.

Making Computerized Adaptive Testing A Diagnostic Tool

✦Hua-Hua Chang, University of Illinois at Urbana-Champaign, 430 Psychology, M/C 716, 630 E Daniel St, Champaign, IL 61821, *hhchang@illinois.edu*

Key Words: sequential design, computerized adaptive testing, cognitive diagnosis, latent trait estimation, latent class classification, educational testing

Computerized adaptive testing (CAT) has become popular in many high-stakes educational testing programs. The goal of a traditional CAT item selection algorithm is to sequentially select items that optimize the estimation process for the examinee's latent trait theta. As a result, only a total score is reported. A new design of CAT also provides feedback about students' individual educational needs in addition to the single overall score. In fact, the US government requires that such diagnostic feedback be provided to parents, teachers, and students as soon as practical. In this paper new methods of classifying examinees into a set of partially ordered latent classes are introduced. The results of both simulation study and a field test of several thousand students will be presented. Issues in estimation accuracy and classification consistency will be addressed. Our research is showing that the proposed Cognitive Diagnostic CAT can be used effectively not only to estimate an examinee's latent trait, but also to classify the examinee's mastery levels of the skills the test is designed to measure. CAT is revolutionarily changing the way we address challenges in assessment and learning.

Flexible Approaches To Cognitive Diagnosis: Nonparametric Methods And Small Sample Techniques

✦Chia-Yi Chiu, Rutgers, The State University of New Jersey, 10 Seminary Place, New Brunswick, NJ 08901, *chia-yi.chiu@gse.rutgers.edu*

Key Words: cognitive diagnosis, Q-matrix

In educational testing research, specialized latent class models for cognitive diagnosis have been developed to classify mastery or non-mastery of each attribute in a set of attributes the exam is designed to assess. The ultimate goal of applying diagnostic models is to classify subjects into one of several different categories describing their attribute profiles. However, in many cases where such models might be applicable, sample sizes are too small to adequately fit models, and even when adequate sample sizes are available, the models often display substantial misfit. An alternative is to use nonparametric methods that only assume a Q-matrix, a matrix that associates items with the specific attributes they measure. Nonparametric methods include cluster analysis techniques and algorithms for classifying examinees based on proximity to ideal response patterns. These robust methods do not rely on fitting models, and result in classification rates that are nearly independent of sample size, and compare favorably to parametric models. These methods and their theoretical properties will be discussed, including rapid algorithms for Q-matrix modification and efficient model-free methods

400 Bayesian modeling and inference for finite populations: A Session in Honor of Joseph Sedransk ■●

Section on Bayesian Statistical Science

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Predictive Inference For Identifying Outliers In Health Care Providers

✦Michael Joseph Racz, Albany College of Pharmacy and Health Sciences, Albany, NY , *michael.racz@acphs.edu*; Guofen Yan, University of Virginia

Key Words: report cards, posterior predictive probability, residual

Provider profiling is the evaluation of the performance of hospitals, doctors, and other medical practitioners to enhance the quality of care. Many jurisdictions have released public report cards comparing hospital or physician-specific outcomes. Given the importance of provider profiling, studying the methodology used and providing enhancements is essential. Ohlssen, Sharples and Spiegelhalter (2007) present an extensive, thoughtful evaluation of "provider profiling" methodology. In particular they are concerned about whether a putative outlier is really an outlier or an observation in the tail of the common distribution for all practitioners, and present methodology to address this issue. We evaluate the Ohlssen et al. (2007) methodology using both NYS bypass surgery data and simulated data of the same type as that used in Racz and Sedransk (2010). We also extend the Ohlssen et al. methodology to permit evaluation of important characteristics seemingly not covered by the Ohlssen et al. procedure. Ohlssen(2007)JRSS A, 170, 865-890. Racz(2010) JASA, 105, 48-58

Bayesian Inference About Variances And Subpopulation Effects In A One-Way Random-Effects Model When The Subpopulations Are Clustered

✦Guofen Yan, University of Virginia, Department of Public Health Sciences, P.O.Box 800717, Charlottesville, VA 22908-0717 USA, *guofen.yan@virginia.edu*

Key Words: Exchangeability, Random effects, Shrinkage, Subpopulation estimates, Survey sample

We provide methodology to relax the assumption that all subpopulation effects in a linear mixed-effects model have, after adjustment for covariates, a common mean. We expand the model specification by assuming that the m subpopulation effects are allowed to cluster into d groups and the composition of the d groups are unknown, a priori. We show that failure to take account of the clustering, as with the customary method, will lead to serious errors in inference about the variances and subpopulation effects, but the proposed model leads to appropriate inferences. The efficacy of the proposed method is evaluated by contrasting it with both the customary method and use of a Dirichlet process prior.

Bayesian Predictive Inference For Finite Population Quantities Under Informative Sampling

✦Junheng Ma, SAMSI, 19 T.W. Alexander Drive, Research Triangle Park, NC 27709, *jma@samsi.info*

Key Words: Bayesian predictive inference, finite population, informative sampling

Bayesian predictive inference is investigated for finite population quantities under informative sampling, i.e., unequal selection probabilities. Only limited information about the sample design is available, i.e., only the first-order selection probabilities corresponding to the sampled units are known. We have developed a full Bayesian approach to make inference for the parameters of the finite population and also predictive inference for the non-sampled units. Thus we can make inference for any characteristic of the finite population quantities. In addition, our methodology, using Markov chain Monte Carlo, avoids the necessity of using asymptotic approximations.

Bayesian Predictive Inference Under Benchmarking Of Body Mass Index And Bone Mineral Density For Small Domains

✦MA CRISELDA TOTO, National Institute of Statistical Sciences, 19 T. W. Alexander Dr, Research Triangle Park, NC 27709, *totomc@niss.org*

Key Words: bayesian predictive inference, multivariate, small area estimation, benchmarking

In sample survey of finite populations, small domains are subpopulations for which the sample sizes are too small for estimation of adequate precision. Considering the population of Mexican American adults (20 years and above) from the large counties of New York, we implement Bayesian predictive inference to estimate the finite population means of body mass index (BMI) and bone mineral density (BMD) from the Third National Health and Nutrition Examination Survey (NHANES

III). Generally, models used in small area estimation do not make use of the unit-level weights. We use a Bayesian nested-error regression model with internal benchmarking constraints that incorporate unit-level sampling weights. Benchmarking is done by applying constraints that will ensure that the `total' of the small domain estimates matches the `grand total'. Benchmarking can help prevent model failure, an important issue in small area estimation. It can also lead to improved prediction because of the information incorporated in the sample space due to the additional constraint. We present results for the multivariate benchmarking Bayesian model and compare the outcomes with its univariate counterpart.

Overview Of Likelihoods For Bayesian Inference About Finite Population Quantities

✦Joe Sedransk, CWRU - Emeritus, 10600 Crossing Creek Road, Potomac, MD 20854, *jxs123@cwru.edu*

Key Words: Prior Distributions, Predictive Inference

Several types of likelihood have been proposed for Bayesian inference about finite population quantities. This talk reviews alternatives and comments about their utilities.

401 Bayesian Models in Life Sciences ●

Section on Bayesian Statistical Science, Section on Health Policy Statistics

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

A Phylogenetic Model Of Language Diversification For The Dating Of Proto-Indo-European

✦Robin Jeremy Ryder, ENSAE, Timbre J340, 3 avenue Pierre Larousse, Malakoff, International 92240 France, *robin.ryder@ensae.fr*; Geoff Keith Nicholls, University of Oxford

Key Words: MCMC, language diversification, dating, phylogenetics

Language diversification is a random process comparable in many ways to biological evolution. We model the diversification of so-called "core" lexical data by a stochastic process on a phylogenetic tree. We focus on the Indo-European language family. The age of the most recent common ancestor of these languages is of particular interest and issues of dating ancient languages have been subject to controversy. We use Markov Chain Monte Carlo to estimate the tree topology, internal node ages and model parameters. Our model includes several aspects specific to language diversification, such as rate heterogeneity and the data registration process, and we show that lexical borrowing does not bias our estimates. We show the robustness of our model and analyse two independent data sets to estimate the age of Proto-Indo-European.

Bayesian Nonparametric Methods For Protein Structure Prediction

✦Kristin Patricia Lennox, Lawrence Livermore National Laboratory, L-229, P.O. Box 808, Livermore, CA 94551-0808, *lennox3@llnl.gov*; David B. Dahl, Department of Statistics, Texas A&M University; Marina Vannucci, Rice University; Ryan Day, University of the Pacific; Jerry W. Tsai, University of the Pacific

Key Words: Dirichlet process mixture model, Bayesian nonparametrics, protein structure prediction, density estimation

The protein structure prediction problem consists of determining a protein's three-dimensional structure from the underlying sequence of amino acids. For template-based modeling, a target is assumed to be structurally similar to other proteins with known structure. We present statistical methods for incorporating information about template protein structures into searches of protein conformation space. The general strategy is to identify a simplified representation of protein structure and then develop a statistical model for nonparametric density estimation. This process is used to first model backbone torsion angles at individual protein sequence positions, then extended to simultaneous modeling at multiple positions. The final modification is to incorporate information about protein side chain positions into the existing backbone model. Analysis of the protein structure data affords the opportunity to explore various extensions to the standard Bayesian density estimation framework, including the incorporation of priors into otherwise nonparametric models and a method for modeling dependence which is an alternative to nonparametric copulas.

Real Time Inference And Risk Prediction For Notifiable Disease Of Animals

✦Chris Jewell, University of Warwick, Department of Statistics, University of Warwick, Coventry, CV4 7AL UK, *chris.jewell@warwick.ac.uk*

Key Words: reversible jump MCMC, epidemic, parallel computing, risk prediction, bayesian, real time

Mathematical modelling for infectious diseases is well established, though obtaining reliable parameter estimates has vexed many attempts at quantitative prediction. Estimates are specific to an outbreak, and must be obtained swiftly in response to an incursion. Case detection lags infection, and censoring of infection events presents a challenging missing data problem. This talk presents a Bayesian data augmentation framework for rigorous inference on SIR-type epidemic models, thereby providing truly quantitative predictions of future risk. A parallelised rjMCMC algorithm is constructed using a heterogeneous population model on individual-level data. Two applications in livestock populations show how the framework can be used to give real-time predictions of, for example, the probability of individual farms becoming infected, the risk farms might pose to the uninfected population if infected, and the locations of infected but undetected farms. This work provides a much-needed real-time decision-support tool for targeting disease control to critical transmission processes, for early detection of infected farms, and for monitoring the efficiency of current control policy.

Hierarchical Models For The Marine Sciences: Analyses Of Climate Variability And Fish Abundance

✦Ricardo Lemos, NOAA/NMFS Environmental Research Division, Southwest Fisheries Science Center, 1352 Lighthouse Avenue, Pacific Grove, CA 93950-2097, *Ricardo.deLemos@noaa.gov*; Bruno Sanso, University of California, Santa Cruz

Key Words: Hierarchical Models, Bayesian approach, Dinamic Linear Models, Discrete Process Convolutions, Parallel computing, Spatiotemporal data analysis

The motivation of this work is the construction of parsimonious models for the marine sciences, making use of available data and understanding of underlying mechanisms. Hierarchical Bayesian methods (HBMs) permit the construction of layered representations of observations, processes and parameters. In HBMs, uncertainty is accommodated explicitly, through the definition of prior distributions for parameters and through the inclusion of error terms. Because model fitting takes place in a single step, estimation uncertainty is properly propagated at all levels. Markov chain Monte Carlo and goodness-of-fit methods are presented. The approach described relies mostly on Discrete Process Convolutions (DPCs) and Dynamic Linear Models (DLMs). Owing to a convolution kernel designed in this work, DPCs are shown to produce adequate spatial interpolations, capturing location-dependent anisotropy and smoothness. DLMs, in turn, allow the breakdown of time-series into multiple components: trends, seasonal cycles and transient fluctuations. Since HBMs can become comprehensive, effort is made to develop efficient algorithms for multiple processing. Large dataset problems are provided as examples.

402 Adaptive Clinical Trial Designs: Advantages over 'Standard' Group Sequential Designs? ■

Biopharmaceutical Section

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Adaptive Clinical Trial Designs: Advantages Over 'Standard' Group Sequential Designs?

✦Cyrus Mehta, Cytel Inc., 675 Massachusetts Avenue, Cambridge, MA 02139, *cyrus@cytel.com*; ✦Scott Emerson, University of Washington, , *semerson@u.washington.edu*; ✦Keaven Anderson, Merck & Co., Inc, 351 N. Sumneytown Pike, UG1C-46, North Wales, PA 19454, *Keaven_Anderson@merck.com*; ✦Roger Lewis, Harbor-UCLA Medical Center, , *roger@emedharbor.edu*

Key Words: Clinical trial, Adaptive design, Group sequential, Operating characteristics

The focus of this panel session will be on issues in clinical trial design, specifically comparing the advantages and disadvantages of "the newer" adaptive clinical trial designs and "the more familiar" group sequential trial designs. Content will be based on a discussion and structured debate amongst the renown leading expert panelists on the ability of each type of trial design (adaptive design vs. 'standard' group sequential design) to adequately address the scientific and statistical issues of randomized clinical trials. Case studies will be used to illustrate issues involved including aspects of sample size modification and population enrichment, two aspects for which adaptive designs are often suggested.

403 In Over Our Heads? Demystifying Complex Problems with Statistical Engineering ■●

Section on Quality and Productivity

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

In Over Our Heads? Demystifying Complex Problems with Statistical Engineering

✦Ronald D Snee, Snee Associates, LLC, 10 Creek Crossing, Newark, DE 19711, *Ron@SneeAssociates.com*; ✦Martha Gardner, General Electric, , *martha.gardner@ge.com*; ✦Erin Tanenbaum, The Nielsen Company, , *Erin.Tanenbaum@nielsen.com*; ✦Will Guthrie, NIST, , *will.guthrie@nist.gov*

Key Words: statistical engineering, applied statistics

The word "complex" is often used in JSM abstracts and is then followed with word such as systems, samples, regression analysis, or the like. In this panel discussion three complex statistical case studies will be reviewed. These problems cannot be found in Chapter 5 of any statistical book and are from various parts of the public sector: market research, technology and manufacturing. Our panelists will focus on how to best utilize statistical theory for practical benefit in solving their complex challenges. A discussion about statistical engineering (the art of applying statistical and probabilistic methods and techniques supporting research and the production of standard reference materials) will solidify the discussion.

404 Gene Expression Normalization and Analysis

Biometrics Section, ENAR

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Gene Identification With True Discovery Rate Degree Of Association Sets And Estimates Corrected For Regression To The Mean

✦Michael Richard Crager, Genomic Health, Inc., 301 Penobscot Drive, Redwood City, CA 94063, *mcrager@genomichealth.com*

Key Words: Gene identification, Interval hypothesis, Regression to the mean, Selection bias, TDRDA set

Analyses for identifying genes with expression associated with some clinical outcome or state are often based on ranked p-values from tests of point null hypotheses of no association. Van de Wiel and Kim take the innovative approach of testing the interval null hypotheses that the degree of association for a gene is less than some value of interest against the alternative that it is greater. Combining this idea with the false discovery rate (FDR) controlling methods of Storey, Taylor and Siegmund gives a computationally simple way to identify true discovery rate degree of association (TDRDA) sets of genes among which a specified proportion are expected to have an absolute association of a specified degree or more. Genes can be ranked using the maximum lower bound (MLB) degree of association for which each gene belongs to a TDRDA set. Estimates of each gene's degree of association

with approximate correction for "selection bias" due to regression to the mean (RM) are derived using simple bivariate normal theory and Efron and Tibshirani's empirical Bayes approach. TDRDA sets, the gene ranking and the RM-corrected estimates of degree of association can be displayed graphically.

Significance Analysis Of Time-Series Gene Expression Profiles :Via Differential/Trajectory Models In Temporal Mrna-Seq Data

✦Sunghee OH, Yale University, 333 cedar street Department of Genetic, School of Medicine Yale University, New haven, CT 06510 USA, *sshshoh1105@gmail.com*; Hongyu Zhao, Yale University; James P. Noonan, Yale University

Key Words: Next generation mRNA-Seq, time-series study, differential and trajectory model, Markov chain Monte Carlo, cross species, semi-parametric and non-parametric

We propose differential and trajectory models for identifying differentially expressed genes over a time course to define a global pattern of gene expression during dynamic transcriptome using next generation mRNA-Seq data application and simulation sets. In times series studies, a variety of statistical models and tools have been developed in capturing and comparing variation in expression over time based on microarray experiment. To the best of our knowledge, this is the first solid study to thoroughly define a statistical trajectory across times points in mRNA-Seq count data with discreteness and over-dispersion property. Semi-parametric, non-parametric methods with pareto and rank approach, and parameter-driven model of stochastic autoregressive poisson model based upon Markov chain Monte Carlo algorithm are developed and applied for time course temporal count data. Specifically, our trajectory approaches for differential analysis are further applied to exploit a comparative analysis in cross-species by identifying significantly species-conserved and species-specific trajectories to investigate dynamic process across species during development.

A Bayesian Hierarchical Model For Correlated Microarray Datasets

✦Bernard Omolo, University of North Carolina at Chapel Hill, 3107F McGavran-Greenberg Hall, CB # 7420, Chapel Hill, NC 27599 USA, *bomolo@bios.unc.edu*; Ming-Hui Chen, University of Connecticut; Haitao Chu, University of Minnesota, School of Public Health, Division of Biostatistics; Joseph G. Ibrahim, University of North Carolina

Key Words: Bayesian hierarchical model, cell-line, correlation coefficient, gene expression, microarray data, probe

Assessment of gene-specific correlation between two independent expression datasets may help in deciding whether to use an original expression data or one updated with additional samples, for differential gene expression analysis. This can be accomplished through modeling the parameters measuring association between variables, for instance, the correlation coefficient. Typically, the correlation coefficients would be computed using the mean expression value for each common gene and cell-line between the two datasets. However, this approach does not utilize the replicated expression values for each gene and instead averages over them, thereby ignoring the effect of multiple probes per gene. We propose a three-level Bayesian hierarchical model for the gene-specific correlation coefficient between two independent datasets that utilizes replicated expression values for each gene. A comparison with the standard approach indicates that the Bayesian approach performs better and hence is more preferable for differential gene expression analysis.

Signals To Be Seen In Tiled Microarray Experimens

✦Sigrun Helga Lund, University of Iceland, Dunhagi 5, Reykjavik, 107 Iceland, *sigrunhelga@gmail.com*; Gunnar Stefansson, University of Iceland, Faculty of Physical Sciences

Key Words: tiled microarray, isothermal probes, repeated probes, false positive, power

The accuracy of signals obtained from tiled microarray experiments has been debated. In this paper four types of signals to be seen in tiled microarray experiments are defined and their replicability tested in three sets of experiments, containing the same samples, but the probesets varying from isothermal probes to probes of equal lengths. The third experiment included repeated copies of each probe which allowed the power of each type of signal to be estimated, both within and between samples. The probability of both false positive and false negative signals for each of the four types was also estimated.

A Comparison Of Methods For Identifying Differentially Expressed Genes In Microarray Experiments

✦Manel Wijesinha, Penn State University, 1031 Edgecomb Avenue, York, PA 17402 USA, *wta@psu.edu*; Dhammika Amaratunga, Johnson & Johnson PRD LLC

Key Words: conditional t, limma, microarray

A common task in microarray experiments is to identify genes that are differentially expressed across two conditions, e.g., in normal tissue vs diseased tissue. This can be addressed quite simply by doing a series of ordinary t-tests. However, several authors have argued that the power of the tests can be improved by borrowing strength across the genes, a process that leads to modified t-tests. Strength can be borrowed either parametrically, leading to methods such as limma, or semi-parametrically, leading to methods such as Conditional t. In this work, we will compare the effectiveness of these methods in situations where standard assumptions, such as normality, hold and in situations where such assumptions do not hold. In addition, we will explore some variations of the volcano plot for gene selection.

A Generalized Linear Model Framework For Underdispersed Count Data With An Application To Mirna Data

✦Lieven Clement, K.U.Leuven, Kapucijnenvoer 35, Leuven, B-3000 Belgium, *lieven.clement@med.kuleuven.be*; Peter Pipelers, Ghent University; Olivier Thas, Ghent University; Jean-Pierre Ottoy, Ghent University

Key Words: Poisson regression, underdispersion, qPCR, GAM, miRNA, Newton-Raphson

miRNA's play an important role in gene regulation and they are often measured by quantitative PCR (qPCR). The output of qPCR is the number of cycles, Cq-value, that is needed for a certain target to exceed

a threshold. The Cq-values can be considered as counts. But, they seem to be underdispersed with respect to the Poisson distribution, i.e. their variance is less than the mean. There exist a rich variety of regression frameworks for overdispersed count data. The choice of alternative distributions for underdispersed count data, however, is rather limited and regression frameworks are lacking. We introduce a tilted Poisson distribution for underdispersed counts. It has two parameters ? and ? for location and scale, respectively. We embed the tilted Poisson distribution in a generalized linear model (GLM) framework. Both parameters ? and ? can be modeled in function of covariates by using linear and/or smooth components. We propose a Newton-Raphson algorithm for parameter estimation and implement it as a new distributional family for the R-package VGAM. We illustrate our approach in a case study for assessing differential expression of miRNA's in the presence of confounders.

Data Preprocessing: Quantification And Normalization Of The Luminex Assay System

✦Eileen Liao, University of California, Los Angeles, Department of Biostatistics, Los Angeles, CA 90095, *biochen@gmail.com*; David Elashoff, University of California, Los Angeles

Key Words: data preprocessing, median normalization, lowess curve, quantile normalization

High throughput genomics experiments generate exhaustive quantities of information. Microarray is one of the technologies that allow rapid analysis of molecular targets of thousands of genes at a time. In microarray experiments, variations in expression measurements emerge from many sources. Despite the huge improvements to the technology, variations still exist. We applied normalization methods to a bead-based multiplex Luminex assay system to reduce the plate-to-plate variability. We quantified performance among measurements of fluorescent intensity, fluorescent intensity minus background, and observed concentration in both high and standard scanning systems, and evaluated the preservation of ranking separation across plates for both standards and control plasma. We then applied scale normalization, lowess curve normalization, smoothing lowess curve extrapolation and quantile normalization to the data, and utilized coefficient of variation across plates to evaluate the performance of each method. Smoothing curve extrapolation is most efficient in reducing variation across plates. Our findings provide some guidance on the selection of normalization methods.

405 Modeling and Testing Using Imaging Data

Biometrics Section, ENAR, International Indian Statistical Association, WNAR

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Assessing Functional Connectivity In Neuronal Networks From Spike-Train Data: A New Class Of Models And Likelihood-Based Inference

✦Kohinoor Dasgupta, University of Michigan Department of Statistics, 439 West Hall, 1085 South University Avenue, Ann Arbor, MI 48109, *kohinoor@umich.edu*; Jijay Nair, University of Michigan; Stilian Stoev, University of Michigan Department of Statistics; Xuanlong Nguyen, University of Michigan Department of Statistics

Key Words: Likelihood based methods, Generalized linear models, Spike-Train data

Recent advances in technology have allowed neuroscientists to collect large amounts of electrophysiological data at fine time scales. An important class of such data is multi-neuronal spike trains - time sequences of firings of a group of neurons. Identifying the functional connectivity of the neurons from spike-train data has been a problem of considerable interest in recent years. This paper proposes a new class of models for characterizing the dependence and connectivity among the neurons over time. Likelihood-based methods for estimating the underlying parameters including the connectivity matrix and base firing rates have also been developed. Asymptotic theory for the estimators is currently under investigation. Our inference methods are shown to be considerably less complex than other comparable methods in the literature and are illustrated on simulation and real data.

Multiscale Adaptive Functional Principal Components Analysis And Its Application In The Neuroimaging Data

✦japing Wang, ENAR, IMS, ASA, 27516 U.S., *jwang@bios.unc.edu*; Hongtu Zhu, University of North Carolina Department of Biostatistics

Key Words: multiscale adpative, functional PCA, neuroimaging data, Adaptive neighborhood, smoothing, kernel estimate

In the functional principle component analysis (fPCA), the raw data often need to be smoothed. The common way is to smooth each observed functional data by either spline or local weighted smoothing method. For example, in neuroimaging data, one assumes voxels are independent then implement the fPCA. However, for the spatio-temporal data, the voxels are spatial dependent. Thus we propose a new approach to find the functional principle components, called multiscale adaptive fPCA. Compared with the existing fPCA, our method has three unique features: being spatial, being hierarchical, and being adaptive. Our approach analyzes all observations in the ellipsoid of each voxel and these consecutively connected ellipsoids across all voxels can capture spatial dependence among imaging observations. This method combines imaging observations with adaptive weights in the voxels within the ellipsoid of the current voxel to adaptively and spatially smooth functional images. Finally, this approach applies the smoothed functional data to the fPCA. Simulation studies and real data analysis are used to demonstrate the methodology and examine the performance of the adaptive estimates.

Determining State Related Changes In Brain Connectivity

✦Ivor Cribben, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, *ijc2104@columbia.edu*; Ragnheidur Helga Haraldsdottir, Columbia University; Tor D Wager, University of Colorado; Martin A Lindquist, Columbia University

Key Words: fMRI, Graph Valued Regression, State-related changes

Most statistical analyses of fMRI data assume that the nature, timing and duration of the psychological processes are known. However, many times it is hard to specify this information a priori. In this work we apply an extension of graph valued regression (Liu, Chen, Lafferty and Wasserman (2010)), a data-driven technique for partitioning the experimental time course into distinct intervals depending on the underlying functional connectivity between certain regions of interest, to describe the changes in brain connectivity that result from a state anxiety induction. The technique builds a tree on the covariate space (time) just as in CART (classification and regression trees), but at each node of the tree estimates a graph, or series of relationships between brain regions. Permutation and bootstrapping methods are performed in order to create meaningful and useful inference procedures. The method is applied to simulated data, as well as, to an fMRI study (n=26) of a state anxiety induction using a social evaluative threat challenge. The results illustrate the methods ability to observe how the networks between different brain regions changed with the subject's emotional state.

Testing Variance Components In A Functional Mixed-Effects Model For Fmri Data

✦Ragnheidur Helga Haraldsdottir, Columbia University, 1255 Amsterdam Ave Rm 1005, 10th Floor, Mail Code: 4690, New York, NY 10027, *ragnheidur@stat.columbia.edu*; Wesley Thompson, University of California, San Diego; Tor D Wager, University of Colorado; Martin A Lindquist, Columbia University

Key Words: fMRI, Brain, FDA, Mixed-effects model

Currently most analysis of multi-subject fMRI data involves fitting a GLM model for each subject, and using the resulting activation parameter estimates in a "second level" group analysis. In this work we introduce a functional mixed-effects model which allows us to not only directly estimate the activation parameters, but also the variance components of the model. Further, we discuss an approach for performing inference on the variance components that allows us to test for significant individual differences between subjects; something not currently performed in fMRI data analysis. We apply the method to the FBIRN data, where a group of subjects perform the same task at 4 different sites.

Model-Based Clustering For Differentiating Lesion Tissue Types In Gadolinium Fmri Scans

✦Roseline Bilina Falafala, Cornell University - Operations Research and Information Engineering, 206 Rhodes Hall, Ithaca, NY 14853, *rb537@cornell.edu*

Key Words: Model-Based clustering, Spatial dependence, Multiple sclerosis

Multiple Sclerosis patients have brain lesions, believed to be caused by a breakdown in the blood-brain barrier (BBB). One way to examine the permeability of the BBB is via injection of Gadolinium (GA), which is visible on a functional Magnetic Resonance Imaging scan. Scans are taken, observing the diffusion of GA into the brain tissue in locations where the BBB is breached. Statistically, we want to identify, characterize, and cluster the lesions of a subject based on the GA diffusion over space and time. For example, GA concentration spikes quickly in young lesions, while older lesions and healthy tissue show very little absorption. We are also interested in the dynamics of GA absorption into different types of lesions. We develop statistical methods for these goals. First, we reduce the data associated with each voxel from a time series to a low-dimensional summary, using regularized functional principal components analysis. Then, we take a spatial model-based clustering approach to cluster the voxels into a small number of different tissue types, taking into account their spatial dependence. Computational methods from image analysis are applied and extended for this purpose

A Residual Bootstrap Method For Evaluating Preprocessing Pipelines In Functional Magnetic Resonance Imaging (Fmri)

✦Xiangxiang Meng, University of Cincinnati, 33 Ridgewood PL, Fort Thomas, KY 41075, *mengxa@mail.uc.edu*; Scott K Holland, Cincinnati Children's Hospital Medical Center; Xiaodong Lin, Rutgers University

Key Words: functional magnetic resonance imaging, bootstrap, motion correciton, general linear model, image reliability

Motion is the major issue that affects the result of the statistical analysis in functional magnetic resonance imaging (fMRI). In this paper, we develop a GLM-based image reliability measure using residual resampling, and evaluate a variety of fMRI preprocessing pipelines for motion correction. Unlike usual cross-validation techniques for imaging reliability, this approach is based on bootstrapping the residuals in the GLM analysis of single-subject fMRI data which does not require homogeneity and independence assumption across subjects. Thus, our approach can be applied to a variety of experimental setting such as longitudinal and case-control. We demonstrate the performance of the proposed image reliability measure using fMRI data corrupted with different levels of motion artifact, and then apply it to evaluate motion correction schemes such as realignment to the first fMRI scan, realignment to the best fMRI scan, and the use of motion parameters as regressors in the GLM analysis.

Joint Modeling Of Mri And Polychotomous Disease Status Using Wavelet With Application To Alzheimer'S Disease

✦Jincao Wu, University of Michigan, 8140-303 Randolph Way, Ellicott City, MD 21043 United States, *jincaowu@umich.edu*; Timothy D. Johnson, University of Michigan

Key Words: Alzheimer's Disease, Prediction, Wavelet, Bayesian Lasso, Polychotomous Data

Our study is motivated by the challenge of using MRI to diagnose Alzheimer's disease. In an MRI study of Alzheimer's patients, white matter changes are highly heterogenous and differ in size and location making it difficult to use MRI as an accurate diagnostic tool. In our study, we propose to jointly model MRI data and polychotomous disease status (normal, mild cognitive impairment or Alzheimer's disease) using wavelets, which can mitigate these problems. In stage I, we apply a 3-D discrete wavelet transformation on the MRI data and shrink the small wavelet coefficients to zero by Bayesian Lasso. In this way, the signal in the data can be represented by a small number of large wavelet coefficients. In stage II, a cumulative probit regression model is used to predict disease status and to select covariates via the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm.

406 Bayesian Modeling for Spatial Data

Section on Bayesian Statistical Science, ENAR, Section on Statistics and the Environment, Section on Quality and Productivity

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Hierarchical Bayesian Analysis Of Directional Data Using The Projected Normal Distribution

✦Fangpo Wang, Department of Statistical Science, Duke University, 214 Old CHemistry Building, Duke University, Durham, NC 27708, *fw19@duke.edu*; Alan E Gelfand, Department of Statistical Science

Key Words: circular data, bimodal distribution, mean direction, MCMC, wave direction

The projected normal distribution is an under-utilized model for explaining directional data. It provides flexibility, e.g., bimodality, asymmetry, and convenient regression specification. We develop fully Bayesian hierarchical models for circular data. We show how they can be fit using suitable latent variables and MCMC methods. We show how posterior estimation of analytical quantities such as mean direction and concentration can be implemented as well as a regression setting. Work in progress will show how we propose to build a projected Gaussian process to capture structured spatial dependence for modeling circular data at different spatial locations. Simulated and real data examples are provided for illustration.

A Statistical Approach To An Ocean Circulation Inverse Problem

✦Seo-eun Choi, Arkansas State University, Department of Mathematics and Statistics, P.O. Box 70, State University, AR 72467 USA, *schoi@astate.edu*; Fred W Huffer, Florida State University; Kevin G Speer, Florida State University

Key Words: MCMC, Bayesian, Inverse problem, Spatial Statistics

It presents, applies, and evaluates a statistical approach to an ocean circulation problem. The objective is to produce a map of ocean velocity in the North Atlantic based on sparse measurements, based on a Bayesian approach with a physical model. The Stommel Gulf Stream model which relates the wind stress curl to the transport stream function is the physical model. A Gibbs sampler is used to extract features from the posterior velocity field. To specify the prior, the equation of the Stommel Gulf Stream model on a two-dimensional grid is used. Comparisons with earlier approaches used by oceanographers are also presented.

On Bayesian "Central Clustering" : Application To Landscape Classification Of Western Ghats

✦Sabyasachi Mukhopadhyay, INDIAN STATISTICAL INSTITUTE, R.A.Fisher Bhavan, 4th Fl., BIRU, 203, B.T.Road, Kolkata, 700108 India, *sabstat123@gmail.com*

Key Words: Cluster analysis, Dirichlet process, Gibbs sampling, Massive data, Mixture Analysis

Landscape classification of the well-known biodiversity hotspot, Western Ghats (mountains), on the west coast of India, is an important part of a worldwide programme of monitoring biodiversity. To this end, a massive vegetation data set, consisting of 51,834 4-variate observations has been clustered into different landscapes by Nagendra and Gadgil (1998). But a study of such importance may be affected by non-uniqueness of cluster analysis and the lack of methods for quantifying uncertainty of the clusterings obtained. Motivated by this applied problem of much scientific importance we propose a new methodology for obtaining the global, as well as the local modes of the posterior distribution of clustering, along with the desired credible and "highest posterior density" regions in a non-parametric Bayesian framework. Clustering of the Western Ghats data using our methods yielded landscape types different from those obtained previously, and provided interesting insights concerning the differences between the results obtained by us and Nagendra and Gadgil (1998).

Bayesian Inference For Complex Computer Models And Large Multivariate Spatial Data For Climate Science

✦K. Sham Bhat, Los Alamos National Laboratory, PO Box 1663, MS-F600, Los Alamos National Laborator, Los Alamos, NM 87545, *bhat9999@lanl.gov*; Murali Haran, Pennsylvania State University; Klaus Keller, Department of Geosciences, Pennsylvania State University Address:

Key Words: computer experiments, hierarchical Bayes, Gaussian processes, multivariate spatial data, uncertainty quantification, climate change

Computer model calibration involves combining information from a complex computer model with physical observations of the process. Computer model output is often in the form of multiple spatial fields, particularly in climate science. We study an effective inferential approach by using Gaussian processes to emulate the computer model, linking the calibration parameters with the multivariate spatial observations. Then we infer the calibration parameters using Bayesian methods, while incorporating more flexible approaches allowing for nonlinear relationships among spatial fields and non-separable covariance functions. In addition, we incorporate the uncertainty due to model discrepancy and measurement error into our inference and predictions, which usually results in more accurate and sharper inference of the calibration parameter and improved characterization of uncertainties. We utilize kernel mixing and matrix identities in order to make computations tractable for large spatial data sets. We apply our approach to infer vertical diffusivity, a climate model parameter from which we obtain projections of the Atlantic Meridional Overturning Circulation (MOC).

Hierarchical Poisson/Gamma Random Field Model

✦Jian Kang, University of Michigan, , *jiankang@umich.edu*; Timothy D. Johnson, University of Michigan; Thomas E. Nichols, University of Warwick

Key Words: Spatial Point Processes, Random Intensity Measure, Classification Model, Nonparametric Bayes, Hierarchical Model

To jointly analyze multiple groups of spatial point patterns, we propose a non-parametric Bayesian modeling approach that extends the Poisson/Gamma random field model (Wolpert and Ickstadt, 1998). In particular, each group of point patterns is modeled as a Poisson point process driven by a random intensity that is a kernel convolution of a

gamma random field. The group-level gamma random fields are linked and modeled as a realization of a common gamma random field shared by all the groups. We resort to a hybrid algorithm with adaptive reject sampling embedded in a Markov chain Monte Carlo algorithm for posterior inference. Also, our model can be used to build a classifier of group label given spatial point patterns based on the corresponding posterior predictive probability. We illustrate our models on simulated examples and two real applications.

Hierarchical Bayesian Random Sets With Applications To Growth Models

✦Athanasios Micheas, University of Missouri-Columbia, Dept of Statistics, MO 65211, *amicheas@stat.missouri.edu*

Key Words: Random Object, Finite Mixture Models, Hierarchical Bayesian Model, Growth Models, Data Augmentation, Reversible Jump MCMC

We propose and study a novel Bayesian Hierarchical framework to model objects stochastically as well as capture the growth or evolution of an object, based on random sets. We efficiently model a random set via a multistage hierarchical Bayesian framework, and using finite mixture models of possibly varying dimension. Such models require the use of data augmentation techniques as well as Reversible Jump MCMC sampling methods. Several growth models are proposed, including Hereditary, Birth-Death and Mixed type, as well as the corresponding Bayesian formulation that provides inference and prediction. The models are applied to modeling forestry (random tree objects) as well as storm cell development as obtained from weather radar over time (storm cell evolution).

407 Robust Methods for Testing, Estimation and Prediction

Biometrics Section, Biopharmaceutical Section, Section on Nonparametric Statistics, Section on Statistics in Epidemiology

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

A Broad Symmetry Criterion for Nonparametric Validity of Parametrically Based Tests in Randomized Trials

✦Russell T Shinohara, Johns Hopkins University, 615 N. Wolfe St. E3033, Baltimore, MD 21231, *taki.shinohara@gmail.com*; Constantine Frangakis, Johns Hopkins University; Constantine Lyketsos, Johns Hopkins University

Key Words: Robustness, Hypothesis Testing, Superefficiency, Randomized Clinical Trial, Causal Inference

Pilot phases of a randomized clinical trial often suggest that a parametric model may be an accurate description of the trial's longitudinal trajectories. However parametric models are often not used for fear that they may invalidate tests of null hypotheses of equality between the experimental groups. Existing work has shown that when, for some types of data, certain parametric models are used, the validity for testing the null is preserved even if the parametric models are incorrect. Here, we provide a broader and easier to check characterization of parametric models that can be used to (a) preserve nonparametric validity of testing the null hypothesis, i.e., even when the models are incorrect, and (b) increase power compared to the non- or semiparametric bounds when the models are close to correct. We demonstrate our results in a clinical trial of depression in Alzheimer's patients.

Challenges of Summarizing Physiological Data from Anesthesia Information Management Systems

✦Michael Bronsert, PhD, Colorado Health Outcomes, University of Colorado Denver, 12477 East 19th Avenue, Mail Stop F443, Aurora, CO 80045, *Michael.Bronsert@UCDenver.edu*; Karl Hammermeister, MD, Colorado Health Outcomes, University of Colorado Denver; William Henderson, PhD, Colorado Health Outcomes, University of Colorado Denver; Michael Mangione, VA Pittsburgh Healthcare System and University of Pittsburgh School of Medicine; Jennifer Nguyen, MD, MEd, Michael E. DeBakey Veterans Affairs Medical Center; John Sum-Ping, VA North Texas Health Care System and UT Southwestern Medical Center; Deyne Bentt, MD, CPHIMS, Washington DC VA Medical Center; David Kazdan, MD, PhD, Cleveland Department of Veteran's Affairs Medical Center; Terri Monk, MD, MS, Durham VA Medical Center

Key Words: Data Reduction, Data Validaion, Summary Measures

Anesthesia information management systems (AIMS) are used to collect perioperative physiologic data. We propose to test the hypothesis that perturbations in perioperative physiologic variables are associated with both short and long term adverse surgical outcomes. We obtained AIMS and surgical data for 19,468 patients having 26,830 surgical operations. The challenges addresed are how or whether to combine disparate field names for some physiologic concepts, removal of outliers or implausible values, and the design of summary measures for data reduction. We made decisions on combining variables on available samples sizes, clinical judgment, and descriptive statistics and distribution of values. Decisions were made about outliers by clinical judgment and large drops in distributional frequencies. Algorithms were developed to eliminate artifactual spikes that are common with invasive monitoring of arterial pressure. We explored data reduction to one or two values per physiologic concept per procedure through the use of area under or over the time-amplitude curve. Examples of our methods will be presented.

Estimating Open Population Site Occupancy Rates From Presence-Absence Data Lacking The Robust Design

✦David Dail, Oregon State University, Department of Statistics, 44 Kidder, Corvallis, OR 97331, *daild@science.oregonstate.edu*; Lisa Madsen, Oregon State University

Key Words: Detection proability, Monitoring, N-mixture models, Open population, site occupancy

It is difficult to obtain accurate estimates of the site occupancy rates for a dynamic animal population when its presence during any survey period is detected with unknown probability. The existing likelihood models that account for unknown detection and allow open populations were proposed for the robust design setting, where several of the survey occasions occur within periods of known population closure. In this paper, we propose an alternative likelihood model that yields an

estimator for the site occupancy rate during every survey period and does not require the robust design. We construct the marginal likelihood of the observed data by conditioning on, and summing out, the actual number of occupied sites during each survey period. A simulation study shows that on average the site occupancy rate estimates from the proposed model are less biased than the estimates from the original likelihood model. Both models are applied to a data set consisting of repeated presence- absence observations of American robins with yearly survey periods, with the proposed model yielding site occupancy rate estimates closer to the estimates obtained from a third model that allows point counts.

Modeling Multivariate Binary Data With Copulas Over Partitions

✦Bruce Judson Swihart, Johns Hopkins Biostatistics, 615 N. Wolfe St., Dept. Biostatistics, Baltimore, MD 21205, *bruce.swihart@gmail.com*; Brian Caffo, Johns Hopkins Department of Biostatistics; Ciprian Crainiceanu, Johns Hopkins University

Key Words: Binary outcomes, Copulas, Marginal likelihood, Multivariate Logit, Multivariate probit, Stable distributions

Many seemingly disparate approaches for marginal modeling have been developed in recent years. We demonstrate that many current approaches for marginal modeling of correlated binary outcomes produce likelihoods that are equivalent to the proposed copula-based models herein. These general copula models of underlying latent threshold random variables yield likelihood-based models for marginal ?xed effects estimation and interpretation in the analysis of correlated binary data with exchangeable correlation structures. Moreover, we propose a nomenclature and set of model relationships that substantially elucidates the complex area of marginalized random intercept models for binary data. A diverse collection of didactic mathematical and numerical examples are given to illustrate concepts.

Partially Monotone Tensor Spline Estimation Of The Joint Distribution Function With Bivariate Current

✦Yuan Wu, University of Iowa, 216 hawkeye court, Iowa City, IA 52246, *yuan-wu@uiowa.edu*

Key Words: Bivariate current status data, Constrained maximum likelihood estimation, Empirical process, Sieve maximum likelihood estimation, Tensor spline basis functions

The analysis of the joint distribution function with bivariate event time data is a challenging problem both theoretically and numerically. This paper develops a tensor spline-based sieve maximum likelihood estimation method to estimate the joint distribution function with bivariate current status data. The I-spline basis functions are used in approximating the joint distribution function in order to simplify the numerical computation of constrained maximum likelihood estimation problem. The generalized gradient projection algorithm is used to compute the constrained optimization problem. The proposed tensor spline-based nonparametric sieve maximum likelihood estimator is shown to be consistent and the rate of convergence can be as good as forth root of sample size under some regularity conditions. The simulation studies with moderate sample sizes are carried out to demonstrate that the finite sample performance of the proposed estimator is generally satisfactory.

Misuse Of Delong Test To Compare Aucs For Nested Models

✦Olga Demler, Boston University, 30 Edge Hill Rd, Newton, MA 02467 USA, *demler@bu.edu*; Michael Pencina, Boston University; Ralph B. D'Agostino, Sr, , Boston University

Key Words: AUC, ROC, risk prediction, DeLong test, logistic regression, model discrimination

Area under the Receiver Operating Characteristics Curve, (AUC of ROC) is a widely used measure of discrimination in risk prediction models. Mann-Whitney statistics is used as a non-parametric estimator of AUC. The difference of two AUCs is often tested by DeLong test. This study was motivated by numerous reports that often the added predictor is statistically significantly associated with the outcome but fails to produce significant improvement in the AUC. We suggest a possible explanation. We show that DeLong test can not be applied to test AUC improvement for nested models for any continuous distribution of the data and very general class of statistical models including logistic regression. First we show empirically that distribution of the difference of two AUCs from nested models is very different from the one used by the DeLong test. We use theory of U-statistics to explain this contradiction by showing that the difference of two AUCs belongs to a degenerate class of U-statistics and therefore has different asymptotic distribution than the one used by the DeLong test. It results in substantial (up to 60%) loss of power by the DeLong test. Possible solutions are discussed.

Nonparametric Covariates Adjustment For Youden Index

✦Haochuan Zhou, Georgia State University, 5234 Spalding Forest CT NE, Atlanta, GA 30328, *mathxzx@langate.gsu.edu*

Key Words: Youden Index, heteroscedastic regression model, local polynomial regression, working sample, asymptotic normality, asymptotic consistency

The receiver characteristic curve (ROC) has been a popular method to evaluate the accuracy of a diagnostic test. Summaries of ROC, the area under the ROC curve and the Youden Index (YI) are two main indexes to measure the capability of a test. Sometimes, the known information from some covariates may influence the accuracy of a test. We propose nonparametric method for covariate adjustment of the YI. Heteroscedastic models under normal and non-normal error assumption are developed and investigated. The local polynomial regression is applied to estimate mean and variance functions of the model. In the model which without normality assumption for the error term, we propose an empirical estimator for YI which efficaciously utilizes available data to construct working samples at any covariate value of interest and is computationally efficient. Under two assumptions, we explore the asymptotic normality, strong uniform convergence rate and asymptotic consistency for the YI estimators. Plenty of simulations are conducted to illustrate the effectiveness for both estimate and the robustness for the empirical estimate. A real data of diabetes disease study is used to demonstrate the methods.

408 Federal Statistics and Methods ■●

Section on Government Statistics, Section on Survey Research Methods, Social Statistics Section, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

The Use Of Longitudinal Analysis To Model Psu Level 1-Year Price Change In The Consumer Price Index

✦John John Schilp, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Washington, DC 20212, *schilp.john@bls.gov*

Key Words: Longitudinal Data Analysis, Consumer Price Index, American Community Survey Data

Longitudinal Data Analysis is used on American Community Survey data to model PSU level 1-year price change in the Consumer Price Index. The American Community Survey (ACS) is a Census product that provides accurate and current demographic estimates every year for statistical agencies uses. Historical modeling methods from previous CPI revisions will be examined while highlighting the benefits of this longitudinal modeling method. Goodness-of-Fit statistics are calculated while Backward Elimination is used to determine a final model. The resulting model will be used to provide variables to stratify Core Based Statistical Areas in the next revision of the Consumer Price Index.

Revising Replicate Selection In The Cpi Variance System

✦Owen J. Shoemaker, Bureau of Labor Statistics, 2 Massachusetts Ave. Rm 3655, Washington, DC 20212, *shoemaker_o@bls.gov*

Key Words: Stratified Random Groups, Replicates

The Consumer Price Index (CPI) program at the Bureau of Labor Statistics (BLS) is actively considering streamlining its CPI Estimation System to produce a more efficient and more flexible overall operation. This New Estimation System will entail the elimination of the replicate structure, which currently provides the necessary replicate values for the CPI's Stratified Random Group (SRG) Variance System. In order for BLS to continue using its SRG methodology, it will become necessary to create these needed replicates "dynamically" (by random assignment) each month from full sample values. In this paper, we will investigate and compare as well as produce the results from the use of dynamically constructed replicate price changes and compare these variance results with the currently computed CPI variances. At least two random methodologies for selecting the replicate values will be analyzed and evaluated. A more robust variance system is the hoped for objective.

Record Linkage Methodology In Longitudinal Database Of Quarterly Census Of Employment And Wages

✦Marek W Kaminski, Bureau of Labor Statistics, 2 Massachusetts Ave NE, Room 4985, Washington, DC 20212-0002 usa, *Kaminski.Mark@bls.gov*; Vinod Kapani, Bureau of Labor Statistics

Key Words: Deterministic Record Linkage, Probabilistic Record Linkage, EM algorithm, Accuracy of Linkage, Missing Data

The Longitudinal Database of Quarterly Census of Employment and Wages (QCEW) links quarterly reports of employment and total wages of all nonfarm establishments covered by State Unemployment Insurance. QCEW is using two types of linkage: deterministic and probabilistic. Deterministic linkage is performed by SESA ID - the combination of state FIPS code, Unemployment Insurance Number, and Reported Unit Number. The remaining part of all establishments, not linked by SESA ID, are then linked by a probabilistic method (weighted matching), based on Fellegi and Sunter theory. In the presented research we perform both SESA ID linkage and weighted linkage. The initial estimate for weights for all variables are computed from frequencies of match and non-match statistics determined by SESA ID match, then the EM algorithm is used for final estimation. The problem of missing data in matching fields is resolved by subdividing the whole data into subsets with complete data for matching fields. In the presented research we evaluate the accuracy and effectiveness of both types of linkage. Recommendations are given.

Variables That Measure The Impact On Collection Rates

Louis Harrell, Bureau of Labor Statistics; Ralf Hertwig, Bureua of of Labor Statisitcs; ✦Jeremy Oreper, Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Washington, DC 20212-0001, *oreper.jeremy@bls.gov*

Key Words: Linear Regression, Durbin-Watson, Data Collection, Current Employment Statistics

The Current Employment Statistics (CES) program produces monthly employment, hours, and earnings estimates for the non-agricultural economy based on a sample of businesses that covers about 410,000 worksites. Estimates are generated using data received from CES respondents over a three month collection cycle. Data collected during the 10 to 16 collection days of each month are used to produce preliminary estimates. This research uses linear regression to measure the impact of the major variables believed to negatively impact the preliminary collection. The indicator variables evaluated include: holidays and vacation time, catastrophic events, reporting method, the size or number of employees per establishment, the length of the payroll period, respondent burden and number of payrolls. These factors may impede reporting measured by the monthly collection rate. The analysis is conducted over a 5-year period, 2006 -2010. Durbin-Watson statistics influence the correct specification of the model and standard statistical tests determine the significance of the regression, coefficients and hence the variables.

Improving Estimates Of Employment In Expanding And Contracting Businesses

Kenneth W Robertson, Bureau of Labor Statistics; Joshua T Duffin, Bureau of Labor Statistics; ✦Jennifer Kim, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Washington, DC 20212, *kim.jennifer@bls.gov*

Key Words: ARIMA-X12, Current Employment Statistics, business employment dynamics, Quarterly Census of Employment and Wages

Kenneth W. Robertson, Joshua Duffin U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Washington, D.C. 20212 The Current Employment Statistics (CES) program produces estimates of employment, hours, and earnings by industry on a monthly basis for the nonagricultural economy. Recent research has suggested that the CES respondent data can be used to produce estimates of employment change in businesses with expanding and contracting employment. These research estimates have included a small overestimate because we have not accounted for the prior month employment of establishments going out of business. In this paper we will describe research to quantify, and to potentially account for, this issue. Data from the Quarterly Census of Employment and Wages program will be used with CES data to assess the size of the overestimate. ARIMA-X12 will be used to develop forecasts of factors that we will use to adjust the estimates to account for this error. The forecast factors will be compared to the actual factors to assess the feasibility of modeling these values.

Who Is Eligible For U.S. Employer-Sponsored Pensions?

✦Patricia Jeanne Fisher, U.S. Bureau of the Census, 4600 Silver Hill Road, Suitland, MD 20746, *Patricia.J.Fisher@census.gov*

Key Words: SIPP, Employer Sponsored Pensions, Retirement

United States (U.S.) private firms as well as state and local governments and the federal government offer employer sponsored pension plans to their employees. In 2006, sixty percent of the U.S.workforce worked for an employer who offered one or more pension plans. Employer sponsored pension plans fall into three categories: a traditional defined benefit plan, an individual contribution plan or a cash balance plan. The U.S. does not require employers to offer pension plans to their employees. U.S. employers set their own conditions on qualification for pension coverage. Results can be used to understand differences between workers who are eligible to participate and those who are not. Results also draws upon the different types of pension plans offered by employers. This paper uses the 2001 and 2004 Panels of the Survey of Income and Program Participation (SIPP) survey. The paper will explain the most common reasons for employees to be ineligible to participate in their employer-sponsored pension plans. Some of the characteristics between employees will be examined. Finally, it will provide an overview of the U.S. workforce retirement employer-sponsored coverage and participation.

Effects Of The Pretrial Supervision Phase On Children Of Defendants: Opportunities For New Data Collection And Analysis

✦Mary McGraw-Gross, StatAid, 6930 Carroll Ave, Suite 420, Takoma Park, MD 20912, *mary@stataid.org*; Cynthia McMurray, National Society of Public Affairs and Administration; Michael Kisielewski, StatAid; Valerie Cruz,

Key Words: data collection, federal statistics, incarceration, survey data, focus groups, data analysis

Research indicates that children of criminal offenders are at higher risk of experiencing adverse social, economic, and developmental effects than their peers. They are six times more likely to be imprisoned, and one-tenth of them will be incarcerated prior to adulthood. Traditionally interventions to mitigate those effects have begun during the post-conviction phase. This underscores the need for data collection and analysis during the pretrial phase to determine whether providing services earlier would yield equal or greater benefits than post-conviction services. This paper outlines children's economic, social, and cultural rights to create a baseline for studying how well those rights are maintained from a parent's pretrial supervision phase to post-incarceration. It then proposes a combined framework of survey and focus group data from agencies or programs to assess the experiences of children during pre-sentencing. Thus, this data collection and analysis will identify beneficial practices-including innovative actions/services-needed to ensure that children's rights are appropriately represented in all stages of the legal process.

409 New Topics in Bayesian Modeling

■●

Section on Statistical Computing, Section on Bayesian Statistical Science, Section on Statistical Graphics, Section on Quality and Productivity

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Modeling Item-Item Similarities for Personalized Recommendations on Yahoo! Front Page

Deepak Agarwal, Yahoo!; ✦Liang Zhang, Yahoo!, 4401 Great America parkway, Santa Clara, CA 95054, *liangzha@yahoo-inc.com*; Rahul Mazumder, Stanford University

Key Words: Collaborative filtering, Item-item similarity, Covariance selection, Bayesian hierarchical models

We consider the problem of algorithmically recommending items to users on a Yahoo! front page module. Our approach is based on a novel multi-level hierarchical model that we refer to as User Profile Model with Graphical Lasso (UPG). The UPG provides personalized recommendation to users by simultaneously incorporating both user covariates and historical user interactions with items in a model based way. In fact, we build a per-item regression model based on a rich set of user covariates and estimate individual user affinity to items by introducing a latent random vector for each user. The vector random effects are assumed to be drawn from a prior with a precision matrix that measures residual partial associations among items. To ensure better estimates of precision matrix in high-dimensional settings, we also impose sparsity through a Lasso penalty on the matrix elements, by taking recourse to the Graphical Lasso algorithm for covariance selection in the M-step. Through extensive experiments on the Yahoo! front page data and the MovieLens data set, we show that our UPG model significantly improves performance compared to several state-of-the-art methods in the literature.

New Approximate Bayesian Confidence Intervals for the Coefficient of Variation of a Gaussian Distribution

✦VINCENT A. R. CAMARA, Research Center for Bayesian Applications Inc., 8799 Bardmoor Blvd., Largo,, FL 33777 USA, *gvcamara@ij.net*

Key Words: Estimation; Loss functions; Confidence Intervals, Statistical analysis.

Abstract: The aim of the present study is to obtain and compare confidence intervals for the coefficient of variation of a Gaussian distribution. Considering the square error and the Higgins-Tsokos loss functions, approximate Bayesian confidence intervals for the coefficient of variation of a normal population are derived. Using normal data and SAS software, the obtained approximate Bayesian confidence intervals will then be compared to the ones obtained by Miller(1991). It is shown that the proposed approximate Bayesian approach relies only on the observations. Miller approach that uses the standard normal distribution does not always yield the best confidence intervals. In fact, the proposed approach has great coverage accuracy and performs often better.

A Bayesian Approach To Non-Crossing Quantile Regression Curve Estimation

✦Yuzhi Cai, University of Plymouth, School of Computing and Mathematics, Plymouth, PL4 8AA UK, *ycai@plymouth.ac.uk*

Key Words: Bayesian method, comonotonicity, MCMC, non-crossing, quantile curves, quantile regression

Conventional estimation methods for quantile regression models do not guarantee the estimated quantile curves to be non-crossing. In this paper a novel Bayesian approach to non-crossing quantile regression curve estimation is proposed. This approach allows us to simultaneously estimate a sequence of non-crossing quantile curves of a response variable conditional on some covariates. It can be used to deal with both time series and independent statistical data. In the simulation studies, we considered two different types of models, one is a time series model, and another is a usual regression model. We also carried out an empirical application to the kidney transplant death data. All the results show that the developed method works very well.

Bayesian Demographer: Software For Probabilistic Population Projections

✦Hana Sevcikova, University of Washington, Seattle, WA , *hanas@uw.edu*; Adrian Raftery, University of Washington; Leontine Alkema, National University of Singapore

Key Words: Bayesian hierarchical model, Markov chain Monte Carlo, fertility projection, life expectancy, population projection, United Nations

Every two years, the United Nations Population Division (UNPD) publishes projections of the populations of all countries of the world, called the World Population Prospects. Instead of assessing uncertainty about the projections, the UNPD produces High, Medium and Low variants of these. However, these are scenarios and lack a probabilistic interpretation. We have now developed new methods for probabilistic projection of the total fertility rate (TFR) and life expectancy for all countries, and the UNPD is planning to use them as inputs to their 2010 projections (to be published in 2011), as a first step towards fully probabilistic projections in the future. To support the UNPD in this effort, we have developed several R packages, that implement probabilistic projections of TFR (bayesTFR) and life expectancy (bayesLife), as well as an easy to use graphical user interface (bayesDem) that combines these components into one convenient bundle. In this talk, a brief background of the projection methods will be given. Then the structure of the software components will be discussed, and the software will be demonstrated.

Modeling Parasite Infection Dynamics When There Is Heterogeneity And Imperfect Detectability

✦Na Cui, University of Illinois at Urbana-Champaign, IL 61821 U.S., *nacui2@uiuc.edu*; Dylan Small, University of Pennsylvania; Yuguo Chen, University of Illinois at Urbana-Champaign

Key Words: Panel data, Infection rate, Recovery rate, Markov chain Monte Carlo, Bayesian hierarchical model

Infection with the parasite Giardia lamblia is a problem among children in Kenya. Understanding the infection and recovery rate is valuable for public health planning. Two challenges in modeling these rates are that infection status is only observed at discrete times even though infection and recovery take place in continuous time and detectability of infection is imperfect. We address these issues through a Bayesian hierarchical model based on a random effects Weibull distribution. The model incorporates heterogeneity of the infection and recovery rate among individuals and allows for imperfect detectability. We estimate the model by a Markov chain Monte Carlo algorithm with data augmentation. We present simulation studies and an application to an infection study about Giardia lamblia.

Bayes Factor For Nonlinear Mixed Effects Model With Dp Prior

✦huaiye zhang, Virginia Tech, 5100G, 1252 progress, Blacksburg, VA 24060 USA, *zhanghy@vt.edu*; huaiye zhang, Virginia Tech

Key Words: Nonlinear mixed effects model, Bayes factor, Dirichlet Process Prior, Marginal likelihood, Longitudinal data

We present a framework for comparing two types of nonlinear mixed effects models, constructed with Dirichlet Process Mixture priors, with alternative parametric Bayesian models. Dirichlet Process Mixture prior is an attractive option to solve longitudinal mixed effects model, but it is not fully discussed under nonlinear model setting. We give the fitting procedure for estimating two types of nonlinear mixed effects model with Dirichlet Process Mixture priors, especially for non-conjugate posterior. Nonlinear mixed effects model is rarely involved in earlier work for Bayes factor. The difficulty is the unknown likelihood constants for some parameters. We note that, given a sample simulated from the posterior distribution, the required marginal likelihood may be simulation-consistently estimated by the harmonic mean of the associated likelihood values; a modification of this estimator that reduces instability is proposed. More work is involved in the framework for estimating the likelihood of Dirichlet Process Mixture model. Simulated and real data involving longitudinal nonlinear mixed effects model are used to illustrate the implementation.

Algebraic Statistics Framework For Causal Inference And Data Privacy With Discrete Data

✦Aleksandra Slavkovic, The Pennsylvania State University, 412 Thomas, University Park, State College, PA 16801, *sesa@psu.edu*; Vishesh Karwa, The Pennsylvania State University

Key Words: Algebraic Statistics, Markov Bases, Causal Inference, Latent Variables, Privacy, Discrete data

We present an algebraic computational framework that handles special cases of latent class analyses. Specifically, we consider discrete data problems with unobserved variables such that arbitrary linear constraints are imposed on the possible realizations of the complete data, and thus on the possible states of the joint distribution of all the variables (observed and unobserved) in the analysis. The constraints are imposed either by the modeling assumptions, the structure of the latent variables or for consistency reasons. We illustrate our methods by applying them to two important related problems. The ?rst problem pertains to the assessment of disclosure risk of releasing potentially sensitive information from a latent class analysis in the form of class membership probabilities and probability distribution of covariates conditional on the classes. The second problem pertains to estimation of average causal effect in presence of unobserved confounders, under the Neyman-Rubin framework of potential outcomes. The code is implemented in R, but interfaces with 4ti2.

410 Splines and Other Penalized Methods ●

Section on Nonparametric Statistics, International Indian Statistical Association

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Radial Basis Function Regularization For Linear Inverse Problems With Random Noise

✦Carlos Valencia, Georgia Institute of Technology, 1017 North Ave NE, Apt B, Atlanta, GA 30306-4414 US, *carlos.valencia@gatech.edu*; Ming Yuan, Georgia Institute of Technology

Key Words: Inverse problems, minimax, regularization, radial basis function

In this paper, we study the statistical properties of method of regularization with radial basis functions in the context of linear inverse problems. Radial basis function regularization is widely used in machine learning because of its demonstrated effectiveness in numerous applications and computational advantages. From a statistical viewpoint, one of the main advantages of radial basis function regularization in general and Gaussian radial basis function regularization in particular is their ability to adapt to varying degree of smoothness in a direct problem. We show here that similar approaches for inverse problems not only share such adaptivity to the smoothness of the signal but also can accommodate different degrees of ill-posedness. These results render further theoretical support to the superior performance observed empirically for radial basis function regularization.

Data-Driven Smoothing Parameter Selection For Estimating Average Treatment Effects

✦Jenny H‰ggstr^m, Department of Statistics, UmeÂ University, UmeÂ, International 90187 Sweden, *jenny.haggstrom@stat.umu.se*; Xavier de Luna, Department of Statistics, UmeÂ University

Key Words: Causal inference, Double smoothing, Local linear regression, Smoothing parameter

The nonparametric estimation of average treatment effects in observational studies relies on controlling for confounding covariates through smoothing regression methods such as kernel, splines or local polynomial regression. Such regression methods are tuned via smoothing parameters which regulates the amount of degrees of freedom used in the fit. In this paper we propose data-driven methods for selecting smoothing parameters when the interest lies in estimating an average treatment effect. For that purpose, we propose to estimate the exact expression of the mean squared error of the estimators. Asymptotic approximations indicate that the smoothing parameters minimizing this mean squared error converges to zero faster than the optimal smoothing parameter for the estimation of the regression functions. In a simulation study we show that the proposed data-driven methods for selecting the smoothing parameters can yield lower empirical mean squared error than other methods available such as, e.g., cross-validation.

Parameter Estimation For Ordinary Differential Equations: An Alternative View On Penalty

✦Yun Li, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109 USA, *yunlisph@umich.edu*

Key Words: Penalty parameter, predator-prey data, profile estimation, smoothing, spline-basis

Dynamic modeling through solving ordinary differential equations has ample applications in many fields. The recently proposed parameter-cascades estimation procedure with a penalized estimation component (Ramsay et al., 2007) combines the strengths of basis-function approximation, profile-based estimation and computation feasibility. Consequently, it has become a very popular estimation procedure. In this paper, we take an alternative view through variance evaluation on the penalized estimation component within the parameter-cascades procedure. We found that the penalty term in the profile component could increase estimation variation. Further, contrary to the traditional belief established from the penalized spline literature, this penalty term in the ordinary differential equations setup also makes the procedure more sensitive to the number of basis functions. By taking the penalty parameter to its limit, we propose an alternative estimation procedure. The simulation studies indicate that our proposed method outperforms the popular penalty-based method, and in real data analysis two methods give similar outcomes. We also provide theoretical properties for the proposed estimator.

Spline Estimation With Shape Constraints

David Papp, Rutgers University; ✦Farid Alizadeh, Rutgers University, Piscataway, NJ 08854, *alizadeh@rutcor.rutgers.edu*

Key Words: spline estimation, shape constraints, optimization

Statistical estimation problems often involve one or several restrictions on the shape of underlying function: monotonicity and convexity constraints, and bounds on the function or its derivatives. A general framework is presented in which most such problems can be jointly investigated, and common numerical methods for finding the optimal estimator can be derived. We shall show that if the estimator is a spline, then many of these problems are tractable (especially in the univariate case) both in the theoretical and practical sense. For the intractable (multivariate) cases tractable approximations are derived. The approach

is illustrated by a variety of applications. Wherever possible, the numerical results are compared to those obtained using methods previously proposed in the literature.

Shape-Restricted Penalized Splines

✦Xiao Wang, Purdue Unviersity, , *wangxiao@purdue.edu*

Key Words: Green's Function, Complementarity Condition, Penalized Splines

Estimation of shape-restricted functions has broad applications in statistics, engineering, and science. In this talk, we first study the shape-restricted penalized spline regression estimators using constrained dynamical optimization techniques. The underlying regression function is approximated by a B-spline of an arbitrary degree subject to an arbitrary order difference penalty. The optimality conditions for spline coefficients give rise to a size-dependent complementarity problem. As a key technical result of the talk, the uniform Lipschitz property of optimal spline coefficients is established by exploiting piecewise linear and polyhedral theory. This property forms a cornerstone for stochastic boundedness, uniform convergence, and boundary consistency of the estimator. The estimator is then approximated by a solution of a differential equation subject to boundary conditions. This allows the estimator to be represented by a kernel regression estimator defined by a related Green's function of an ODE. The asymptotic normality is established at interior points via the Green's function.

Bivariate Penalized Splines

✦Luo Xiao, Cornell University, 101 Malott Hall, Dept. of Statistical Science, Cornell Univ., Ithaca, NY 14850, *lx42@cornell.edu*; Yingxing Li, Cornell University; David Ruppert, Cornell University

Key Words: Asymptotics, Bivariate Smoothing, Nonparametric Regression, Penalized Splines, Thin Plate Splines, Covariance Function

We propose a new penalized spline method for bivariate smoothing. Tensor product B-splines with row and column penalties are used as in the bivariate P-spline of Marx and Eilers (2005). What is new here is the introduction of a third penalty term and a modification of the row and column penalties. We call the new estimator a Bivariate Penalized Spline or BPS. The modified penalty used by the BPS results in considerable simplifications that speed computations and facilitate asymptotic analysis. We derive a central limit theorem for the BPS, with simple expressions for the asymptotic bias and variance, by showing that the BPS is asymptotically equivalent to a bivariate kernel regression estimator with a product kernel. As far as we are aware, this is the first central limit theorem for a bivariate spline estimator of any type. We also derive a fast algorithm for the BPS. Our simulation study shows that the mean square error of the BPS is comparable to or smaller than that of other methods for bivariate spline smoothing. Examples are given to illustrate the BPS.

Smoothing Mechanism Of Cyclic Cubic Regression Splines Smoothing

✦Mihoko Minami, Keio University, 3-14-1 Hiyoshi Kohoku-ku, Yokohama, International 223-8522 JAPAN, *mminami@math.keio.ac.jp*

Key Words: Hat matrix, eigen decomposition, penalized likelihood, periodic variation, harmonic functions

Cyclic cubic regression spline smoothing is a method to estimate s periodic smooth regression function such as daily or annual pattern of temperature. For a given set of knots, a cyclic cubic spline function is a piecewise cubic function continuous up to second derivatives at the knots, and the function values and derivatives up to the second order at the both endpoints are equal. When the knots are equally spaced, the vector of function values holds an eqation with the vector f second derivatives and cyclic band matrices. From this equation, it is shown that the eigen values and vectors of the hat matrix for cyclic cubic regression spline smoothing have an interesting structure and it characterizes smoothing mechanism of cyclic cubic regression spline smoothing.

411 Quantile Regression ●

Section on Nonparametric Statistics, International Indian Statistical Association, Section on Statistical Computing

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Nonparametric Estimation Of A Censored Additive Quantile Regression

✦Dawit Zerom, California State University at Fullerton, CA 92831, *dzerom@fullerton.edu*; Ali Gannoun, University of Montpellier II

Key Words: Censored, additive, nonparametric, quantile

A two-step nonparametric procedure is proposed for estimating the additive components of a censored additive quantile regression model. Asymptotic properties are established. To evaluate its numerical performance, a simulation study is also provided.

Model Diagnostics for Quantile Regression

✦Guixian Lin, R&D at SAS Institute Inc, 100 SAS campus Dr., CARY, NC 57513, *guixian.lin@sas.com*; Huixia Judy Wang, North Carolina State University

Key Words: Quantile regression, Model misspecification, Goodness of fit, Inference of quantile regression process, Empirical process, Residuals plots

Quantile regression is emerging as an attractive alternative method to the ordinary least squares regression, and has been widely used in practice. It is very powerful in analyzing data with heterogeneity, and is particularly useful in detecting evolving covariate effects along the quantiles. For example, in medical studies, we might be interested in detecting heterogeneous treatment effects concentrated in the upper tails. To make valid and efficient inferences in such situations, we need to test model adequacy for a rang of quantiles instead of a particular quantile. Here we propose a test based on the cumulative sums process of the residuals. The two key advantages of the proposed method are: 1. nonparametric smoothing is not involved; 2. plots of the cumulative residual processes can identify the nature of model misspecification. The asymptotic properties of of the test statistic are derived. Simulation results demonstrate the satisfactory performance of the proposed test in finite sample size according to its power and size.

Nonparametric Quantile Regression with Heavy-Tailed and Strongly Dependent Errors

✦Toshio Honda, Hitotsubashi University, Graduate School of Economics, Hitotsubashi Univ., 2-1 Naka, Kunitachi, Tokyo, 186-8601 JAPAN, *honda@econ.hit-u.ac.jp*

Key Words: conditional quantile, random design, long-rande dependence, stable distribution, martingale CLT, linear process

We consider nonparametric estimation of the conditional q-th quantile for stationary time series. We deal with stationary time series with strong time dependence and heavy tails under the setting of random design. We estimate the conditional q-th quantile by local linear regression and investigate the asymptotic properties. It is shown that the asymptotic properties are affected by both the time dependence and the tail index of the errors. The results of a small simulation study are also given.

Self-Normalized Based Approach To Nonparametric Inference

Zhibiao Zhao, Penn State University; Xiaofeng Shao, University of Illinois at Urbana-Champaign; ✦Seonjin Kim, Penn State University, 326 Thomas Building, University Park, PA 16802, *szk172@psu.edu*

Key Words: Nonparametric regression, Self-normalization, Quantile regression, Conditional heteroscedasticity

In nonparametric inference problems, limiting variance function in asymptotic normal distributions often depends on specific model structure and admits a complicated form that may contain unknown nonparametric functions. Traditional approaches construct consistent estimate of the limiting variance function through extra smoothing procedure, which may deliver very unstable results. In this article, we propose self-normalization based approaches to address nonparametric inference problems without estimating the limiting variance functions. It is shown that the new approach has several advantages over the traditional ones. Monte Carlo simulations are conducted to compare the finite sample performance of the self-normalization based approaches with the traditional ones. Illustrations using real data examples are presented.

A Bayesian Approach To Multiple Quantiles Estimation In Regression

✦Yunwen Yang, University of Illinois at Urbana-Champaign, 725 South Wright Street, 114 Illini Hall, Champaign, IL 61820, *andrea.yang2@gmail.com*; Xuming He, University of Illinois at Urbana-Champaign

Key Words: Quantile regression, multiple quantiles, empirical likelihood, Bayesian

Quantile regression has developed into a systematic methodology for estimation of conditional quantile functions. Usually, quantile regression estimation is carried out at one percentile level at a time, and the resulting estimates tend to have high variability in the data sparse areas. In this talk, we consider a Bayesian empirical likelihood approach (BEL) to quantile regression, which can naturally incorporate various forms of informative priors to explore the commonality across quantiles. The focus of the presentation is to show how the BEL approach facilitates an efficient way of joint estimation of several quantiles, leading to more efficient quantile estimation, especially in the data sparse areas. We show that the posterior-based inference for BEL is asymptotically valid, and demonstrate both theoretically and empirically how the BEL approach improves efficiency over the usual quantile regression estimators. Finally, we use the BEL approach to quantile regression as a statistical downscaling method in climate studies, and illustrate by example the merit of our proposed BEL approach.

Solving Hinge-Loss Support Vector Machine With Quantile Regression

✦Yonggang Yao, SAS Institute Inc., 100 SAS Campus Drive, Cary, NC 27513, *yonggang.yao@sas.com*; Guixian Lin, R&D at SAS Institute Inc

Key Words: Support Vector Machine, Quantile Regression, Rank Score Test, Likelihood Ratio Test

Hinge-loss for support vector machine (SVM) and check-loss for quantile regression (QR) are the most popular L1 loss functions that have been widely applied for respectively solving classification and regression problems. In this report, we describe how a hinge-loss SVM problem can be cast as a check-loss QR problem. We also numerically investigate several QR inference methods in order to solving SVM problems, and our simulation studys show that rank score and likelihood ratio methods are of good performance to make statistical inference for SVM classification methods.

Power Transformed Quanitle Regression With A Limited Dependent Variable

✦Hyokyoung Grace Hong, Baruch college, CUNY, , *hyokyoung.hong@baruch.cuny.edu*

Key Words: power transformation, quantile regression, Medical expense, skewed data

Econometric modeling of healthcare costs or medical expenses involves an analysis of heavily skewed data with a limited dependent variable, which is continuous over most of its distribution but has a mass of observations at one or †more specific values, such as zero. The skewness in outcome can be dealt with the power transformation. Here we propose†the estimation method of the conditional quantiles based on the censored and power transformed quantile regressions. The quantiles are estimated using the location-scale model, which facilitates simpler analysis owing to reduction in the number of parameters used and enables more intuitive interpretation. The usefulness of the proposed model will be demonstrated with the 2007 medical expenditure panel survey data.

412 Advancing the Frontier of Statistical Consulting Using Innovative Measurement, Methods and Models

Section on Statistical Consulting, Section on Quality and Productivity

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

A Guide To The Design And Analysis Of Small Clinical Studies

✦Farid Kianifard, Novartis Pharmaceuticals, One Health Plaza, East Hanover, NJ 07936, *farid.kianifard@novartis.com*; M. Zahur Islam, Novartis Pharmaceuticals

Key Words: exploratory study, pilot study, proof of concept, sample size, clinical trial

Clinical studies which have a small number of patients are conducted by pharmaceutical companies and research institutions. Examples of constraints which lead to a small clinical study include a single investigative site with a highly specialized expertise or equipment, rare diseases, and limited time and budget. We consider the following topics which we believe will be helpful for the investigator and statistician working together on the design and analysis of small clinical studies: definitions of various types of small studies (exploratory, pilot, proof of concept); bias and ways to mitigate the bias; commonly used study designs for randomized and nonrandomized studies, and some less commonly used designs; potential ethical issues associated with small underpowered clinical studies; sample size for small studies; statistical analysis methods for different types of variables and multiplicity issues. We conclude the paper with recommendations made by an Institute of Medicine committee, which was asked to assess the current methodologies and appropriate situations for conducting small clinical studies. (To appear in Pharmaceutical Statistics, 2011)

Impact Of Randomization And Concurrent Controls In A Pilot Study Designed To Provide Preliminary Evidence To Support Subsequent Investigation

✦Rickey E Carter, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, *carter.rickey@mayo.edu*; Qian Shi, Mayo Clinic

Key Words: Pilot study, Study design

A well-designed pilot study helps advance science by providing essential preliminary data to support or motivate further study. Often, preliminary evidence of efficacy is sought after to provide rationale for the continuation of the investigations of a candidate intervention. When an intervention's efficacy is quantified as a Bernoulli random variable, for example, tumor response in oncology studies, probability mass functions can be enumerated to determine the probability that the observed result from a pilot study supports further evaluation of the intervention. In this paper, an 'efficacy signal' is computed using one- and two-sample pilot study designs. Efficacy signal is defined as the probability of observing a more favorable response proportion relative to a historical control (one sample setting) or the probability of having the new intervention's response proportion numerically superior to a concurrent control (two sample setting). In this sense, the 'efficacy signal' can be viewed as an innovative power function for pilot studies. Recommendations for pilot study designs will be drawn from case studies emanating from our Center for Translational Science Activities (CTSA).

Survival Analysis and Model Comparisons for a Breast Cancer Data Set

✦Hong Li, Cameron University, Lawton, OK 73505, *lhong@cameron.edu*

Key Words: Proportional hazards model, Median survival time, Survival function, Hazard function, Risk factors

A survival analysis on a data set of 295 early breast cancer patients from the Netherlands Cancer Institute is performed in this paper. We assume a proportional hazards model, and select two sets of risk factors for death and metastasis for breast cancer patients respectively by using standard variable selection methods. In addition, a newly developed probability distribution-hypertabastic distribution and the hypertabastic proportional hazards model are introduced. To evaluate the performance of the new model and compare it with other popular distributions, we fit the Cox, Weibull and log-logistic models, in addition to the hypertabastic model. The results from the model comparisons show that the hypertabastic proportional hazards model outperforms all the comparison models when fitted to the breast cancer data. It indicates that it can be a flexible and promising alternative to the practitioners in this field.

Direct Inference Requiring Only A Touch Of Bayes: An Alternative To The Nissen-Wolski Meta Analysis Questioning The Safety Of Avandia (Rosiglitazone)

✦Ralph G. O'Brien, Case Western Reserve University, Department of Epidemiology and Biostatistics, Cleveland, OH 44122, *obrienralph@gmail.com*

In raising concerns about the safety of Avandia, the meta analyses by Nissen and Wolski (2007, 2010) helped precipitate the 9/2010 FDA decision to markedly restrict the drug's use. However, the N&W methodology drew sharp criticism, and analyses by others reached different conclusions. Here, the NW2010 data is used to illustrate a proposed inference method, and its associated point estimate and interval. Let theta be the true difference in event proportions, p.Avandia - p.comparator. Consider the three hypothesis intervals, {-} = {theta < -0.0005}; a null interval, {0} = {-0.0005 <= theta <= 0.0005}; and {+} = {theta > 0.0005}. Let the a priori priors be p{-} = 0.025, p{0} = 0.95, p{+} = 0.025. Using NW2010, the posteriors are p{+ | data} = 0.014 for death and 0.809 for MI. Consider the point null, {.} = {theta = theta0}. The theta0 that maximizes the cumulative Bayes factor for {.}, provides a point estimate; here, 0.0000 for death and 0.0001 for MI. 9:1 support intervals come from finding delta such that p{0} = p{theta0 - delta <= theta <= theta0 + delta} = 0.10 goes to p{0 | data} = 0.90. For NW2010, these intervals are [-0.003, 0.003] for death and [-0.004, 0.005] for MI.

Use Of Zero-Inflated Mixture Models To Compare Antibody Titers Among Asthma Sub-Populations In Response To The H1N1 Vaccine

✦Leela M Aertker, RHO, 6330 Quadrangle Dr, Suite 500, Chapel Hill, NC 27517, *Leela_Aertker@RhoWorld.com*; Daniel J Zaccaro, RHO

Key Words: vaccine, zero-inflated model, asthma, antibodies, influenza, H1N1

Pandemic H1N1 vaccine was administered to participants with mild/moderate and severe asthma. H1N1 antibody titers were measured pre-vaccination (Day 1) and post-vaccination (Day 21). A preponderance of titers below the lower detection limit was observed (36% - 75% of observations at Day 1, depending on subgroup). Titers above the upper detection limit (0% - 8% at Day 21) were also observed. Zero-

inflated log-normal models that accounted for left- and right-censoring and a "point mass" were utilized to compare subgroups with respect to antibody titers. Results derived from traditional analytical methods such as imputation of censored values were compared to results from zero-inflated methods. By formal criteria, zero-inflated models provided a better fit to data and yielded results that were qualitatively and quantitatively different from traditional models. Zero-inflated models yielded geometric mean titers that were from 2- to 3-fold different from traditional models and elucidated differences among subgroups that were obscured by traditional methods. The apparent reduced immunogenicity of vaccine among previous (seasonal) vaccine recipients was investigated.

"Knock Knock": Walk-In With A Quantitative Questionnaire

✦Mehary T. Stafford, University of North Texas/University of Texas at Austin, Dallas, TX 75231 United States, *drmehary@yahoo.com*

Key Words: survey, method, walk-in, response, rate, quantitative

This study explains an effective method of collecting data from a group of scientists working in a region with a historic low response rate. Researchers seek to get relevant data to answer specific research questions. One of the major challenges is getting enough data to analyze. In a time of information overload there is still a need to collect more data from people groups we know little about. However, collecting data from a region with a historic low response rate is a major challenge. This study used a walk-in method of distributing questionnaires in person. Four hundred questionnaires were distributed to a university faculty and 298 questionnaires were returned resulting in a 74.5 % response rate. After exclusion of 12 cases with missing information, 286 cases (71.5 % response rate) were analyzed. Most of the respondents were men (M=92.1%, F=7.9 %).

413 Advances in Inference and Estimation Methodology

Section on Survey Research Methods

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Pre-Sampling Model-Based Inference III

✦Stephen Woodruff, Specified Designs, 800 West View Terrace, Alexandria, VA 22301-2750, *stevewoodruff@comcast.net*

Key Words: Model Based Inference, Design Based Inference, Pre-sampling Inference

In survey sampling, a sample unit`s study variables are expanded to population totals by probability design based (DB) or model based (MB) expansions that implicitly treat a unit`s study variables as totals over entities called atoms contained in a unit. Expansions would make little sense if applied to unit statistics other than atom totals, for example, industry surveys where units are businesses, atoms are employees, and unit study variables are establishment totals for hours, wages, and number of workers. Other examples are household, mail, and ecological surveys. Woodruff (2010, 2009), derived Pre-sampling Model Based Inference from this atom structure, a structure that depends on probability sampling of population units and of atoms that comprise each unit. It provides estimates that retain the best properties of both MB and DB inference and that eliminate the main shortcomings of each. The result can be order-of-magnitude error reduction. Sampling error under repeated sampling from stratified cluster designs is the basis for comparison of the Combined Ratio Estimator and the Pre-sampling Model Based Estimator. Formulae for sampling errors are derived and analyzed.

Rule Of Thumb Regarding The Use Weights In Survey Sampling Analyses

✦Marnie Bertolet, University of Pittsburgh, Department of Epidemiology/EDC, 127 Parran Hall/130 Desoto St., Pittsburgh, PA 15217, *mhb12@pitt.edu*

Key Words: Survey Sampling, Inverse Probability Weights, Survey Weights, Multi-Level Models, Linear Mixed-Effects Models

When using model-based inference in survey sampling, many practitioners are unsure as to whether sampling weights should be incorporated into the analysis. Weighted estimates can protect against informative sampling bias, however they have a larger variance that increases null results. A common rule of thumb is to compare the unweighted and weighted analyses. If the estimates match then the analysis is "correct" and you can use the more optimal unweighted analysis. If the estimates do not match, then you have non-ignorable design or model misspecification. In this case, more care is needed in model building. If a model cannot be found where the weighted and unweighted analyses match, then the weighted analysis should be used. While this advice is easy to give and implement, it is not always accurate. Using a set of simulations on survey weighted mixed-effect models; I provide counter-examples to this advice. I also provide guidance on when this advice may be correct, and when it may not be correct.

Hierarchical Design-Based Estimation In Stratified Multipurpose Surveys

✦Hugo Andres Gutierrez, Universidad Santo Tomas, Av caracas No 49 55 apt 216, Bogot·, International Colombia, *psirusteam@gmail.com*

Key Words: Design-based estimation, finite population, hierarchical estimation, multipurpose surveys, stratified sampling

This paper considers the joint estimation of population totals for different variables of interest in multi-purpose surveys that make use of stratified sampling designs. When the finite population has a hierarchical structure, different methods of unbiased estimation are proposed. Based on Monte Carlo simulations, it is concluded that the proposed approach is better, in terms of relative efficiency, than other suitable methods such as the generalized weight share method.

Robust Estimation In Two-Phase Sampling

✦David Haziza, University of Montreal, CP 6128 Succ. Centre ville, Montreal, QC H3C3J7 Canada, *david.haziza@umontreal.ca*

Key Words: Inluential unit, robust estimation, two-phase sampling

Influential units, also called outliers, often occur in practice, especially in business surveys due to the skewness of the distribution of economic variables collected by this type of surveys. A unit is said to be influential when its inclusion or exclusion from the sample has an important

impact on the magnitude of survey statistics (e.g., estimated totals). In this presentation, we consider the problem of robust estimation in the context of two-phase sampling designs. In this context, the total error of an estimator can be expressed as the sum of the error due to the first phase and that due to the second phase. An influential unit may potentially have an impact on both errors. We define the concept of conditional bias attached to a unit with respect to both phases and show it can be viewed as a measure of its contribution to the total error. Following Beaumont, Haziza and Ruiz-Gazen (2010), we propose a class of robust estimators for two-phase sampling. In the presence of unit nonresponse, the set of respondents is often viewed as a second phase of selection. The proposed method can thus be naturally extended to that case. Results of a limited simulation study will be shown.

A Coverage Approach To Evaluating Mean Square Error

✦alan h dorfman, bls, 7807 custer rd, bethesda, MD 20814, *dorfman_a@bls.gov*

We propose a new method for evaluating the mean square error (mse) of a possibly biased estimator, or, rather, the class of estimators to which it belongs. The method uses confidence intervals c of a corresponding unbiased estimator and makes its assessment based on the extent to which c includes the (possibly biased) estimator of interest. The method does not require an estimate, implicit or explicit, of the bias of the estimator of interest, is indifferent to the bias/variance breakdown of its mse, and does not require surety of the model on which it is based.

Approximate Confidence Intervals For A Parameter Of The Negative Hypergeometric Distribution

Lei Zhang, Mississippi State Department of Health; ✦William D Johnson, Pennington Biomedical Research Center, Baton Rouge, LA , *William.Johnson@pbrc.edu*

Key Words: Confidence interval, Empirical coverage probability, Inverse sampling, Large sample, Negative hypergeometric distribution

The negative hypergeometric distribution (NHD) is of interest in applications of inverse sampling without replacement from a finite population where a binary observation is made on each sampling unit. Thus, sampling is performed by randomly choosing units sequentially one at a time until a specified number of one of the two types is selected. Assuming the total number of units in the population is known but the number of each type is not, we investigate the maximum likelihood estimator (MLE) and an unbiased estimator for the unknown parameter. We use Taylor's series to develop five approximations for the variance of the parameter estimators. We then propose five large sample confidence intervals (CIs) for the parameter. Based on these results, we simulated a large number of samples from various NHDs to investigate performance in terms of empirical probability of parameter coverage and CI length. The unbiased estimator is a better point estimator relative to the MLE as evidenced by empirical estimates of closeness to the true parameter. CIs of the unbiased estimator tended to be shorter because of its relatively small variance but at a slight cost in terms of coverage probability.

Resampling Variance Estimation For Two-Phase Sample

✦Hyunshik Lee, Westat, 1600 Research Blvd, Rockville, MD 20850, *hyunshiklee@westat.com*; David A Marker, Westat

Key Words: Double-expansion estimate, reweighted expansion estimate, jackknife, bootstrap, Taylor method, relative performance

Two-phase sampling is often used in a wide variety of surveys. Variance estimation from a two-phase sample has been a subject of active research. The re-sampling method of variance estimation has been used for this problem. However, the method confronts a challenging problem when the first phase sampling fraction is high. In the extreme (but not uncommon) case some first-phase strata are take-all, but there is subsampling at the second phase. This issue is studied theoretically and using simulation.

414 Bayesian Analysis in Clinical Trials ■

Biopharmaceutical Section, Section on Bayesian Statistical Science
Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Analysis of Drug Effect on Counts of Lymphocyte Subsets

✦Jiacheng Yuan, Astellas Pharma Global Development, Inc., 3 Parkway North, Deerfield, IL 60015 USA, *jason.yuan@us.astellas.com*

Key Words: Bayesian method, clinical study, CD antigen, MCMC, total lymphocyte count

A Bayesian analysis method with MCMC sampling is suggested to analyze the counts of subsets of the total lymphocytes collected in clinical studies using the software WinBugs. Two major assumptions are made: (1) the count of lymphocytes with a set of antigens is the product of the total lymphocyte count (TLC), the set of proportions corresponding to lymphocytes with every individual antigens, and a multiplier for the interaction among the set of antigens under consideration; and (2) the drug effect impacts the TLC, as well as the proportions of cells with specific antigens. The analysis model also accounts for the positive correlation of changes from baseline of the same endpoint between two post-dose time points.

An Approach To Bayesian Sample Sizing

Robb Muirhead, Statistical Consulting; ✦Adina Soaita, Pfizer Inc, 102 Lovers Lane, East Lyme, CT 06333 USA, *adinasoaita@gmail.com*

Key Words: Bayesian methods, smple size determinationa, clinical trials

There is growing interest in the pharmaceutical industry in incorporating Bayesian design and analysis concepts in clinical trials in all phases of drug development. The focus of this talk is on the design aspect, and in particular on sample size determination (SSD). We will focus specifically on a "proper Bayesian" approach for SDD, in which only Bayesian concepts are used in both the design and analysis stages.

Using Bayesian Method To Predict Responders Based On Clinical Scores In The Initial Weeks Of Treatment In A Phase Ii Multiple Sclerosis Trial

✦Dazhe Wang, sanofi-aventis Biostatistics & Programming, 9 Great Valley Parkway, Malvern, PA 19425 USA, *dazhe.wang@sanofi-aventis.com*; Lynn Wei, Sanofi-Aventis

Key Words: Bayesian predictive model, Responders prediction, ROC curve analysis

In the treatment for Multiple Sclerosis (MS), responders and non-responders to study medication may have different profiles for the clinical outcome of interest, and the separation of the profiles may start at the early stage of the treatment due to a possible early onset of drug action. The ability to predict responders based on early observations from patients can help make timely decision to stop treatment for those who are predicted as non-responders. The aim of this research was to develop Bayesian prognostic models for allocation of a patient to either responder or non-responder group at study end on the basis of partial data collected at the initial stage of the study. Three independent models (logistic, longitudinal and autoregressive) were considered using the data from a phase II MS clinical trial. Their predictive performance was then evaluated using the receiver operating characteristic (ROC) curve analysis.

A Dose-Finding Design For Combination Therapy With Bayesian Bivariate Ordinal Probit Model

✦Rui Qin, Mayo Clinic, 200 First St SW, Rochester, MN 55905, *qin.rui@mayo.edu*; Jianchang Lin, Florida State University; Sumithra J. Mandrekar, Mayo Clinic; Daniel J Sargent, Mayo Clinic

Key Words: dose finding, bivariate ordinal probit model, continual reassessment method, combination therapy

Combination therapies are becoming increasingly popular in oncology for potential synergistic efficacy. While previous knowledge about each individual drug may suggest a proper dosing range for combination therapy, a separate dose-finding trial is generally required to determine the optimal dose combination. We propose a dose-finding design with Bayesian bivariate ordinal probit model incorporating toxicity and efficacy as ordinal variables. The contribution of each agent in the combination therapy is modeled in the linear predictors of ordinal outcomes. Priors are used to capture historical information of single agents and the Monte Carlo Markov Chain approach is used for parameter estimation and guiding dose-escalation. Simulation studies are conducted to understand and assess the operating characteristics of this design under multiple scenarios. Considering 50% as a benchmark for the recommendation rate of the optimal dose region (within 15% of the pre-specified optimal dose level), our design performs well under most scenarios considered (most between 50.2%-65.8%, except two scenarios with 32.4%, 43.2%).

A Risk Assessment Framework For Decision Making In Drug Product Development

✦Amit Phansalkar, Straight Line Performance Solutions, 43 Bennington St, Newton, MA 02458 USA, *amit@cognika.com*

Key Words: Risk Assessment, Bayesian modeling, scenario analysis tool, quantitative decision making

The issues surrounding low productivity and escalating costs of drug product development have been well documented in the past. The overriding concern with making product portfolio decisions is lack of information. The ability to quantify risks and identify the influence of behavioral bias towards project decisions is crucial to the successful launch of a drug product. We present a Bayesian framework (tool) for modeling uncertainty and quantitative decision making for drug product development. The quantitative risk evaluation method is implemented as a scenario analysis tool estimating the probability of success of a decision at various stages of drug development. It also helps ferret out possible biases and assumptions in the critical factors and compare alternatives. The tool identifies factors that are likely predictive in the selection of most viable alternative in making the decision. Based on the set of initial assumptions regarding the influence of these factors, the tool estimates the percentage risk associated with each decision. A case study is presented to illustrate the modeling procedure and the effectiveness of the tool in making project selection decision.

A Transparent And Efficient Approach In Proof Of Concept Study Design And Subsequent Decision Making

✦Gang Jia, Merck and Company, 1517 Wynnemoor Way, Fort Washington, PA 19034 USA, *gajst72002@yahoo.com*; Paul Delucca, Merck and Company; Bruce Binkowitz, Merck & Co Inc; Daniel Bloomfield, Merck

Key Words: Clinical trial, proof of concept, Sample size, Bayesian analysis, Decision analysis, Study design

Consider designing a two-arm parallel POC (proof of concept) clinical study in which the primary endpoint is a continuous and normally distributed variable. The goal of the POC study is to obtain information on the efficacy and safety of the experiment drug and to make reasonable decision in the next step of development. Generally resources are limited at this stage of the development process, which poses the question of how to decide on the size of the POC study to adequately serve the purpose of the study. The usual approach of sample size calculation based on Type I error rate and study power is confusing and doesn't provide the relevant information for the purpose of the study. Here we provide a framework of the study design and analysis based on a transparent decision making process. We also propose to incorporate Bayesian analysis approaches to have the flexibility to apply all the relevant information in the decision making.

415 Causal Inference and Estimation

Social Statistics Section

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Bias Correction In Small Area Estimation When Covariates Are Measured With Error

✦Trijya Singh, Texas A&M University, Department of Statistics, College Station, 601 Cross Street, Apt. # 26, College Station, TX 77840 U.S.A., *trijyas@stat.tamu.edu*

Key Words: Fay-Herriot model, Corrected Scores, Additive measurement error

In this paper we discuss the corrected scores approach to deal with measurement error in covariates in the small area estimation problem. This approach was originally suggested by Nakamura (1987) in various kinds of measurement error models and we study its applicability and efficiency in the Fay-Herriot small area estimation model when the auxiliary information is measured with error.

Non-Parametric Weighting Methods For Estimating Mediation Effects: An Application To The National Evaluation Of Welfare-To-Work Strategies

✦Guanglei Hong, University of Chicago, 1200 East Madison Park, Unit 2, Chicago, IL 60615, *ghong@uchicago.edu*; Jonah Deutsch, University of Chicago; Heather Hill, University of Chicago

Key Words: Causal inference, Controlled direct effect, Mediation, Natural direct and indirect effects, Non-parametric, Propensity score

Labor force participation has become necessary for maintaining welfare eligibility since 1997. Experimental evidence suggests that mandating work may increase maternal depression among some sub-groups. This study investigates employment as a potential mediator of the policy effect on depression by analyzing the experimental data from the National Evaluation of Welfare-to-Work Strategies. We use the marginal mean weighting through stratification (MMW-S) method to approximate a two-stage randomized experiment-participants are randomly assigned to either the Labor Force Attachment (LFA) program or the control condition and are subsequently assigned to either employment or unemployment. This non-parametric approach to bias reduction estimates with robustness the controlled direct effect of the policy and that of employment. Moreover, by combining the MMW-S method with the ratio-of-mediator-probability weighting (RMPW) method, we estimate the natural direct and indirect effects of the policy. The paper clarifies the statistical assumptions, delineates the analytic procedure, and presents the empirical results. We assess the performance of the algorithm through a simulation study.

Extending The Baron And Kenny Mediation Analysis To Allow For Exposure-Mediator Interactions: Sas And Spss Macros

✦Linda Valeri, Harvard, 665 Huntington Avenue, Boston, MA 02115, *lvaleri@hsph.harvard.edu*; Tyler J VanderWeele, Harvard University

Key Words: Mediation, Interaction, Software

Mediation analysis is a useful and widely employed approach to studies in the field of psychology and in the social and biomedical sciences. The main contribution of the present work is providing SAS and SPSS macros for easy implementation of mediation analysis allowing for the presence of exposure-mediator interactions. The software enables an investigator to carry out mediation analysis under a regression framework, building on the traditional approach proposed by Baron and Kenny (1986). This traditional approach does not allow for the presence of exposure-mediator interaction in the mediation analysis model and our aim is to fill this gap. We describe identifiability conditions and counterfactual definitions for the causal effects estimated via mediation analysis with a particular emphasis on the case in which exposure-mediator interaction is present. This is accomplished by applying and extending the work on identification and estimation of direct and indirect causal effects of VanderWeele and Vansteelandt (2009,2010).

Individual Change Models For The Analysis Of Randomized Longitudinal Designs: The Role Of The Pretest

✦Joseph R Rausch, Cincinnati Children's Hospital Medical Center,College of Medicine,University of Cincinnati, 3333 Burnet Ave, Cincinnati, OH 45230, *joseph.rausch@cchmc.org*

Key Words: randomized, longitudinal, pretest, individual change, clinical trial, multilevel models

Randomized longitudinal designs provide a number of advantages over randomized pre-post designs, including the ability to employ individual change models for the examination of treatment effects. In this context, individual change models allow for a more precise specification of the treatment effect and increased power and precision. However, the appropriate role of the pretest within such models has generally been unclear in the literature. The present talk provides clarity on this issue, including how the models of interest can be fit within a multilevel modeling framework. Furthermore, the models of interest are compared with respect to statistical validity and efficiency in complete and missing data scenarios. It is concluded that researchers should generally use (a) the pretest as part of the outcome vector and (b) some form of the pretest as a covariate when using individual change models to examine treatment effects in randomized longitudinal designs.

Interval Matching: Propensity Score Matching Using Case-Specific Bootstrap Confidence Intervals

✦Wei Pan, University of Cincinnati, P.O. Box 210049, Cincinnati, OH 45221-0049, *wei.pan@uc.edu*; Haiyan Bai, University of Central Florida

Key Words: propensity score matching, propensity score analysis, caliper matching, bootstrap

Propensity score matching is an essential procedure in propensity score analysis. The current procedure in propensity score matching is to match each of the treated cases with one or more control cases based on closest propensity scores which are the point estimates of the likelihood of the cases to be assigned into the treatment group. The problem with this procedure is that it is difficult to establish a criterion to evaluate the closeness of the matched cases without knowing the standard error of the estimate of each case's propensity score. Cochran and Rubin (1973) suggested using a caliper band to avoid "bad" matches. However, this fixed or case-invariant caliper band still cannot address the standard errors of the estimates of propensity scores. The present study proposes interval matching to capture the standard error of the estimate of the propensity score for each case using case-specific bootstrap confidence intervals. In this proposed interval matching, if the confidence interval of a treated case overlaps with that of one or more control cases, they will be taken as "good" matches. The implementation of interval matching is illustrated with an empirical example.

Propensity Score Matching In Time-To-Event Analysis

Bo Lu, The Ohio State University; Zhenchao Qian, The Ohio State University; Anna Cunningham, The Ohio State University; ✦Chih-Lin Li, The Ohio State University, , *li.698@buckeyemail.osu.edu*

Key Words: Propensity score matching, Survival data, Sensitivity analysis, Observational study, Causal inference

Cox proportional hazard (PH) model has been widely used to analyze time-to-event data. However, it may produce bias when the proportional hazard assumption is violated or any time-dependent confounders exist. Much research has been focused on using propensity score weighting to control for time-dependent confounders and information censoring, thus substantially reduces biases in estimating effects. In this talk, we consider propensity score matching to balance covariates and incorporate it with Cox PH model. We extend the conventional binary treatment matching into multiple group matching. After matching, we propose a simulation-aided sensitivity analysis to assess the robustness of the estimates to any potential unobserved confounders. The method will be applied to a data to study the causal relation between premarital cohabitation and marital disruption.

Italian Universities Careers: A Competing-Risks Analysis

✦Matilde Bini, Department of Economics - European University of Rome -, V. degli Aldobrandeschi, Rome, International Italy, *bini@ds.unifi.it*; Bruno Monastero, Politecnico di Torino; Margherita Velucchi, Department of Statistics - University of Florence _

Key Words: competing risks analysis, survival analysis, university careers

University teachers are a fundamental component for the research and teaching activities development. In modern societies, it is extremely important to adopt strategic policies of recruitment, promotion and retirement of the academic personnel to improve the turn-over, to satisfy the needs of planning of research and teaching activities, as well as to improve the efficiency of the related economic activities. In this work, we propose to analyze determinants that affect careers promotions of the Italian academic staff from 1988 to 2008 academic years. The study uses a generalized survival analysis (competing risks analysis) that allows to consider the promotion as the cause of failure, and teachers can be considered to be exposed to more than one cause of failure. Information such as sex, age, length of service, job position, academic branch and salaries, universities of teachers are collected in data set provided from Ministry of Research and Italian Consortium of Universities (CINECA).

416 Issues in Adaptive Clinical Trials (II)

Biopharmaceutical Section

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Adaptive Patient Population Selection Design In Clinical Trials

Hui Quan, Sanofi-Aventis; ✦Dongli Zhou, Forest Research Institute, 180 Summit Ave Apt H2, Summit, NJ 07901 U.S., *dongli.zhou@frx.com*; Pierre Mancini, Sanofi-Aventis; Gary G Koch, Department of Biostatistics, Gillings Schools of Global Public Health, UNC

Key Words: conditional power, sample size re-estimation, closed testing procedure, weighted combination test, trial simulation

For the success of development for a new drug, it is crucial to select the sensitive patient populations. To potentially reduce timeline and cost, we may apply an adaptive patient population selection design to a therapeutic trial. In such a design, based on early results of the trial, patient population(s) will be selected/determined for the final stage and analysis. Because of this adaptive nature and the multiple between-treatment comparisons for multiple populations, an alpha adjustment is necessary. In this paper, we propose a closed testing procedure for assessing treatment effects for multiple populations and a weighted combination test for population and sample size adaptations. Computation/simulation is used to compare the performances of the proposed method and the other multiplicity adjustment methods. A trial simulation is presented to illustrate the application of the methods.

Design And Execution Of A 2-Stage Adaptive Design In Pediatric Migraine

✦Christopher Assaid, Merck & Co., 351 North Sumneytown Pike, UG 1C-46, North Wales, PA 19454-2505, *christopher_assaid@merck.com*; Ying Zhang, Merck & Co.; Xiaoyin Frank Fan, Vertex Pharma; Tony W Ho, Merck & Co.

Key Words: Adaptive Design, Placebo Effect, Migraine, Pediatric Clinical Trials

Treatments that have been demonstrated to be effective in reducing/eliminating migraine-related pain and associated symptoms have historically failed to demonstrate efficacy in pediatric patients. One observation from completed trials is that the placebo response rate tends to be substantially higher in pediatric trials as compared to adult trials, leading to the potential for a ceiling effect. In this presentation we will describe the multitude of historical challenges and potential theories that might explain the null results of these trials. We will also describe aspects of a novel adaptive design that were identified with a goal of maximizing the probability of success of a trial of a well-established, effective (in adults) treatment of acute migraine in pediatric patients. Additional discussion will surround some of the logistical issues associated with an adaptive design in acute migraine, particularly around multiple randomizations in a short time window for pediatric patients.

Assessing The Performance Of Adaptive Design Trials Based On Real Clinical Trial Data

✦Suvajit Samanta, Merck Research Laboratories, , *suvajit_samanta@merck.com*; Weili He, Merck & Co., Inc.; Yiyun Tang, Pfizer Inv

Key Words: Adaptive design, clinical trials, simulation methodology, nonparametric bootstrap, permutation method

Adaptive clinical trials hold the promise of drastically improving the clinical development process by optimizing the risk and benefit profiles for patients. The theory and potential benefits of adaptive trials have extensively been discussed in literature. However, little information has been presented on case studies of adaptive design scenarios and the potential gains in efficiency as compared to traditional trials. On that regard, simulations play an important role in assessing the performance of adaptive designs as it is often too complex to determine operating characteristics of adaptive designs, such as power, type I error, number of required patients in simple formulas. The simulation studies make many assumptions including treatment response curve, distribution of responses which are generally unknown at the beginning of clinical program or trial. This research assesses the performance of one adaptive design by simulating the adaptive design based on data from completed real clinical trial using permutation method or nonparametric bootstrap method. The proposed assessment is robust as it does not make any underlying assumptions with regard to treatment response curve.

Confidence Intervals And Point Estimates For Adaptive Group Sequential Trials

✦Lingyun Liu, Cytel Inc, 675 Massachusetts Avenue, Cambridge, MA 02139, *lingyun.liu@cytel.com*; Cyrus Mehta, Cytel Inc.; Ping Gao, The Medicines Company; Pralay Senchaudhuri, Cytel Inc.; Pranab Ghosh, Cytel Inc.

Key Words: Adaptive design, median unbiased estimate, confidence interval, stagewise ordering, flexible design, conditional type I error

Adaptive sequential designs have been intensively investigated in the literature. It is well known that type I error can be preserved by preserving the conditional type I error. The inference problem was addressed by Mehta et al (2007). This approach (RCI), however, is only guaranteed to provide conservative coverage of the treatment effect. In addition, this method cannot produce an unbiased point estimate. Brannath et al (2009) generalizes the stage wise adjusted confidence intervals (SWACI) of Tsiatis et al (1984) to adaptive setting. This method provides nearly exact coverage. Both of these two methods are implemented in EastÆ. The SWACI method is limited to one-sided test and is only applicable when there is a single adaptive change through the whole trial. For one-sided test, the SWACI method can only provide either lower or upper confidence limits but not both at the same time. We offer another approach which provides exact coverage and can be applied to a trial with multiple adaptive changes. Both confidence limits can be obtained using this new approach.

A Recent Case Study Of Operational Phase Ii/Iii Seamless Design

✦Lynn Wei, Sanofi-Aventis, 200 Crossing Blvd, P.O.Box 6890, Bridgewater, NJ 08807, *lynn.wei@sanofi-aventis.com*; Hui QUan, sanofi-aventis; Loic Darchy, sanofi-aventis

Key Words: Power, inferential seamless design, multiplicity adjustment, dose selection

In this presentation, a case study is used to illustrate the application of an operational Phase II/III seamless design. The study consists of two parts but uses one protocol. Four doses of an experimental drug are included in the first part and up to two doses will be selected for the second part based on results from the first part. Data from the first part will not be included in the final analysis and multiplicity adjustment is performed considering only the selected dose(s). We prove the strong controls of the Type I error rate with such an approach even though the dose(s) for the second part cannot be pre-specified at the design stage and any one or two of the four doses can be selected for the second part. We compare such a design with an inferential seamless design which combines data from the two parts in the final analysis but considers all of the four doses in multiplicity adjustment. The operational seamless design is more efficient and requires smaller sample size for achieving the same power since a larger significance level is used. In conclusion, the operational seamless design is more preferable because of its flexibility and simplicity.

417 Macroeconomic Modeling ■

Business and Economic Statistics Section, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Moments to Remember: Some Popular Models for the Distribution of Income

James B. McDonald, Brigham Young University; ✦Jeffrey T. Sorensen, Brigham Young University, 153 FOB, Brigham Young University, Provo, UT 84602, *jtsorensen@gmail.com*; Patrick A. Turley, Harvard University

Key Words: skewness, kurtosis, generalized beta type 2 distribution, generalized gamma distribution

This paper explores the ability of some popular income distributions to model observed skewness and kurtosis. We present the generalized beta type 1 (GB1) and type 2 (GB2) distributions' skewness-kurtosis spaces and clarify and expand on previously known results on other distributions' skewness-kurtosis spaces. Data from the Luxembourg Income Study are used to estimate sample moments and explore the ability of the generalized gamma, Dagum, Singh-Maddala, beta of the first kind, beta of the second kind, GB1, and GB2 distributions to accommodate the skewness and kurtosis values. The GB2 has the flexibility to accurately describe the observed skewness and kurtosis.

A General Procedure For Selecting A Best Multinomial Population With Application

✦Saad T Bakir, Alabama State University, 2405 Reston Place, Montgomery, AL 36117, *sbakir@alasu.edu*

Key Words: Income mobility, ranking and selection, singular multivariate normal

A procedure is developed for selecting a subset which is asserted to contain the "best" of several multinomial populations with a pre-assigned probability of correct selection.According to a pre-chosen linear combination of the multinomial cell probabilities, the "best" population is defined to be the one with the highest such linear combination. The procedure is based on the large sample approximation of the multinomial distribution by the singular multivariate normal distribution. The proposed procedure is applied to real economics data relating to population income mobility in four countries.

Consumer Inflation Expectations In Turkey

✦ECE ORAL, Research and Monetary Policy Department,Central Bank of the Republic of Turkey, Istiklal Caddesi No: 10 Ulus, ANKARA, 06100 TURKEY, *Ece.Oral@tcmb.gov.tr*

Key Words: Consumers, Inflation Expectations, Survey Data, Quantification Methods

The expectations obtained from surveys play an important role as leading indicators for the application of the monetary policies. The ability to measure inflation expectations is an integral part of central bank policy especially for central banks that are implementing inflation-targeting regime. A forward-looking perspective is essential to the success of inflation targeting. Therefore, a central bank having primary objective of price stability are interested in inflation expectations. Qualitative data on inflation expectations obtained from surveys can be quantified into numerical indicators of the expected rates of price change. This paper presents the results of different quantification methods such as Carlson-Parkin method, balance method, regression method put into action in order to estimate Turkish consumer inflation predictions based on monthly consumer surveys. Carlson-Parkin method quantifies qualitative survey data on expectations assuming aggregate expectations are normally distributed. In order to capture non-normality, stable distributions are also considered. The quantification techniques are compared with each other as well.

A Statistical Approach To A Dynamic Flow Model Of The Labor Market

✦Mingfei Li, Bentley University, 175 Forest Street, Waltham, MA 02452, *mli@bentley.edu*; Mihaela Predescu, Bentley University

Key Words: Labor market, business statistics, nonlinear system, parameter estimation

Neugart (2002) proposed a dynamic flow model of the labor market with nonlinear and endogenous outflow rate from unemployment. Inflation as a feedback on job offers through the real money supply was also involved in this model system. And in this unemployment and inflation deterministic model system, it is shown that the stability of "equilibrium of rate of unemployment" affected by the parameters of this system. However, how these parameters should be determined in this system has not been discussed. In this talk we would like to show that for practical purpose, determination of the parameters in this dynamic model is critical, i.e. bad choice of parameters may lead the whole model fail in describing or forecasting unemployment and inflation. We start our stochastic approach from simply including random error terms in the deterministic model, and apply MLE method on the parameters of the system. Simulations on real labor market will show the performance of the model with statistical estimated parameters in predicting real labor market. Furthermore, we discuss an adjusted model by release endogenous inflow rate of labor market.

Bayesian Multivariate Ordinal Probit Analysis Of China'S Consumer Confidence Index Data

✦Junni Zhang, Peking University, , *zjunni@gmail.com*; Wei Lan, Guanghua School of Management, Peking University

Key Words: Multivariate Ordinal Probit Regression, Parameter Expansion, Inverse Wishart Distribution

We build a Bayesian multivariate ordinal probit model, using the parameter expansion approach to expand the correlation matrix into a covariance matrix, and then adopting a specific conjugate prior that can help shrink the off-diagonal elements of the inverse covariance matrix (therefore those of the inverse correlation matrix) toward zero. Gibbs sampling is then used to sample the parameters. We apply this model to analyze China's consumer confidence index data for year 2009, in which consumers' judgements about seven aspects of the economy were quarterly collected for the present status and the status three months later.

A Generalized Maximum Entropy Approach For Estimating Armington Elasticity: An Application For Italy

✦LUCA SECONDI, University of Tuscia - Faculty of Economics, Via del Paradiso 47, Viterbo, International 01100 VITERBO, *secondi@unitus.it*; GUIDO FERRARI, University of Florence & Renmin University of China

Key Words: Armington elasticity, Imperfect Substitutability, CGE model, Generalized Maximum Entropy

Armington assumption enables researchers to differentiate commodities according to place of origin through the introduction of an imperfect substitutability between domestically produced and imported products. This hypothesis and the related elasticity play an essential role in applied modelling such as Computable General Equilibrium (CGE) models. However, due to the lack or paucity of adequate data, Armington elasticities are never estimated, neither directly, nor indirectly for the country or region for which the model is constructed; instead, these behavioural parameters are taken from literature and imputed. In this paper, based on macroeconomic data, we estimate Armington elasticities for Italy by using the Generalized Maximum Entropy (GME) method, so to obtain values that can be regarded as a general frame when elasticities of substitution between domestic and imported goods within branches are needed in General Equilibrium (GE) models computation for developed countries or regions, thus making the imputation consistently macro economically based.

Assessing Monetary and Fiscal Policy on Macroeconomic Fundamentals in Nigeria: Aftermath of the Recent Global Financial Crisis

✦Chioma Nwosu, Central Bank of Nigeria, Central Business District, Abuja, International 00234 Nigeria, *cpnwosu@yahoo.com*; Phebian Omanukwue, Central Bank of Nigeria

Key Words: Monetary Policy, Fiscal Policy, Output, Prices, Finacial Crisis

The paper analyses the effects of fiscal and monetary policy in Nigeria. This debate has resurfaced in the literature with the outturn of the global financial crisis. Thus, this paper seeks to assess policy shocks on fluctuations in output and prices especially during the current crisis period. The analysis further captures the contribution of these shocks to these two real variables. In carrying out the analysis, a Variance Auto-Regressive model (VAR) was developed; first, by analyzing the monetary policy shocks; second, the fiscal policy shocks is then introduced to the model to jointly analyze the effects of the two policy on output and prices.

418 Bioinformatics and Statistical Genetics ■●

IMS, Section on Statistics in Epidemiology, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Comparison Of Tests For Association With Interaction Of Genes Between Two Loci

✦Jie Kate Hu, University of Washington, University of Washington, Department of Biostatistics,F-600, Health Sciences, Seattle, WA 98195-7232, *hujie0704@gmail.com*; Xianlong Wang, Fred Hutchinson Cancer Research Center; Pei Wang, Fred Hutchinson Cancer Research Center

Key Words: Gene-Gene Interaction, GWAS, Linkage Disequilibrium (LD) Test

Detection of gene-gene interaction has become a hot topic in current Genome Wide Association Studies (GWAS). Besides traditional logistic regression analysis, new methods have been developed in recent years such as comparing linkage disequilibrium (LD) in case and control groups, logistic kernel machine based test, and case-only analysis. Our interest is to compare these tests on their powers to detect either interaction effect between two genes at two unlinked loci or the overall association including both interaction and main effect. Previous comparison studies have been confused by quite different and usually unstated definition of interaction. To address this problem, we point out the differences of the assumption and definition for each gene-gene interaction test. We then investigate about ten different tests under different scenarios with the purpose to provide investigators some guidance on choosing appropriate statistical tests for various research interests and background.

A New Method To Compare The Haplotype Distributions Between Populations

✦Liping Tong, Loyola Unversity Chicago, 1032 W Sheridan Rd, Department of Mathematics and Statistics, Chicago, IL 60660, *ltong@luc.edu*

Key Words: haplotype distribution, average mismatch score, quadratic form of multinomial random variables, association study, fixation index

Accurate characterization of haplotype structure and diversity is a key challenge in statistical genetics. We propose a new statistic to assess and compare the haplotype variations among populations which is particularly suited to this emerging challenge. We first describe the properties and calculations of this method. Subsequently, using simulation studies, we show that the proposed method is more powerful than the chi-square test statistic when comparing haplotype distributions under the following two circumstances (1) when variations of haplotype distances are not balanced (2) when haplotypes are tainted by accumulated mutations or genotyping errors. We also performed simulations to show that the proposed method can be applied to case-control association study and can be much more efficient than the single locus association test by greatly decreasing number of multiple tests. Finally, we applied our method to the Human Genome Diversity Project (HGDP) and HapMap3 data for SNPs on chromosome 2 in the region surrounding the LCT gene. Our results showed that 726 pairs of populations (out of 780) can be distinguished (p-value < 0.05) using the 127 SNPs surrounding the LCT gene.

On Rogers'S Proof Of Identifiability For The Gtr+Gamma+I Model

✦Juanjuan Chai, Department of Mathematics, Indiana University, Rawles Hall, 831 E 3rd Street, Bloomington, IN 47405, *chaij@indiana.edu*; Elizabeth Ann Housworth, Department of Mathematics, Biology and Statistics, Indiana University

Key Words: Identifiability, Gamma distribution, invariable sites, general time reversible Markov model

Recently, Allman etc. pointed out an error in Rogers's proof of the identifiability of the popular general time reversible Markov model for DNA evolution with heterogeneous rates coming from a mixture of a Gamma distribution and invariable sites in phylogenetics. We provide a proof for the claim under dispute in Rogers's paper and thus complete the proof of generic identifiability for this model using only pairwise comparisons with calculus technique. Rate matrices with only one non-zero eigenvalue and phylogenies with only one or two distinct pairwise inter-species distances form the basis of the exceptional cases. However, we can identify when the collection of pairwise joint sequence distributions comes from these exceptional cases. It is not currently known whether using three-taxon or higher-order comparisons can yield identifiability for the GTR+Gamma+I model in these exceptional cases.

Copy Number Variation Detection Using Next Generation Sequencing Data

✦Heng Wang, Iowa State University, , *hengw@iastate.edu*; Dan Nettleton, Iowa State University

Key Words: copy number variation, next generation sequencing, Hidden Markov Model

Modern genomic technologies enable us to understand the genetic mechanisms of cancers and diseases. Recently developed next generation sequencing technology shows advantages in the aspects of resolution and accuracy in detecting disease related copy number variations (CNV). We present a novel algorithm for CNV detection using next generation sequencing data. Our method incorporates Hidden Markov Models and Bayesian methods to obtain the posterior probability of each underlying copy number "state" for every aligned position along the genome. The proposed method also accounts for spatial dependencies among positions. We use simulation and analysis of real data sets to compare our method with recently published approaches.

Mathematically-Based General Framework For Integrating Multiple Heterogeneous Existing Data Sets Into A Novel Data Collection

✦Jozsef Bukszar, Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University, 1112 East Clay Street, Richmond, VA 23298, *jbukszar@vcu.edu*; Amit Khachane, Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University; Karolina Aberg, Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University; Youfang Liu, Department of Genetics, University of North Carolina at Chapel Hill; Joseph McClay,

Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University; Patrick Sullivan , Department of Genetics, University of North Carolina at Chapel Hill; Edwin van den Oord, Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University

Key Words: data integration, genetics, heterogeneous data, multiple-hypothesis testing, applied mathematics

We present a general framework that integrates information from multiple heterogeneous existing data sets (EDS) into a novel data collection (NDC) in order to find genetic units related to a certain disease. The framework is general in the sense that the EDS-s can be of any type whose genetic units (SNP-s, genes, chromosomal regions) can be ranked, where ties are allowed. The NDC may be of any type that has a statistic value assigned to each of its genetic unit, e.g. next-generation sequencing data or GWAS. The methods of the framework rely on exact mathematical formulas that ultimately provide the posterior probability that a genetic unit in the NDC has an effect. While the formulas rely on numerous unknown parameters, almost all of these unknown parameters can be aggregated into a single term, which, therefore, can be estimated precisely. The framework also includes some other tools, such as methods testing EDS for being informative to the NDC, and assessing the increase in power of adding another EDS. The methods have been validated through simulations, cross-validations on multiple schizophrenia GWAS, and an actual replication study.

An Empirical Bayesian Model For Identifying Differentially Co-Expressed Genes

✦John Alexander Dawson, Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Room 4720A, Madison, WI 53706, *jadawson@wisc.edu*; Christina Kendziorski, Departments of Statistics and of Biostatistics, University of Wisconsin-Madison

Key Words: Empirical Bayes, Differential correlation, Gene Expression, Microarray, Meta-analysis

A common goal of high-throughput genomic experiments is to identify genes that vary across biological conditions. Most often this is accomplished by identifying differentially expressed genes and many effective methods have been developed for this task. Although useful, these approaches do not accommodate other types of differential regulation, such as differential co-expression (DC). Investigations of DC genes are hampered by large search-space cardinality and outliers and as a result, existing DC approaches are often underpowered, prone to false discoveries, and/or computationally intractable for even moderately sized datasets. To address this, an empirical Bayesian approach is developed for identifying DC gene pairs within a single study or across multiple studies. The approach provides a false discovery rate (FDR) controlled list of significant DC gene pairs without sacrificing power. Computational complexity is eased by a modification of the EM algorithm and procedural heuristics. Simulations suggest that the approach outperforms existing methods; and case study results demonstrate utility of the approach in practice.

Gene Ontology-Based Over-Representation Analysis Using A Bayesian Approach

✦Jing Cao, Southern Methodist University, TX 75206, *jcao@smu.edu*

Key Words: over-representation analysis, gene ontology, high-throughput experiment, Bayesian model

In high-throughput experiments a common strategy to interpret the results is to detect a list of differentially expressed genes and then to use the knowledge of the functional characteristics of the genes as a means to gain insights into the biological mechanisms underlying the gene list. Specifically, over-representation analysis (ORA) is conducted to investigate whether gene sets associated with particular biological functions, for example as represented by Gene Ontology (GO) annotations, are over-represented in the gene list. However, the standard ORA analyzes each GO term in isolation and does not take into account the dependence structure of the GO term hierarchy. We have developed a Bayesian approach to measure over-representation of GO terms that incorporates the GO dependence structure by taking into account evidence not only from individual GO terms, but also from their related terms. The Bayesian framework borrows information across related GO terms to strengthen the detection of over-representation signals. As a result, this method tends to identify biological pathways associated with subtrees of interacting GO terms rather than individual isolated GO terms.

419 Assessment in the Classroom ■●

Section on Statistical Education, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Classroom Experiment Comparing Student Lab Activities To Instructor Demonstrations

✦Karen McGaughey, Cal Poly, 1 Grand Avenue, Statistics Department, San Luis Obispo, CA 93407, *kmcgaugh@calpoly.edu*; Soma Roy, Cal Poly; Beth Chance, California Polytechnic State University; Alan Rossman, California Polytechnic State University - San Luis Obispo

Key Words: classroom research, java applets, active learning, teaching statistics

The use of technology is ubiquitous in today's statistics courses. Analysis software such as Minitab and R is used to provide data analysis experience. Java applets are used to develop conceptual understanding of means and medians, sampling distributions and p-values. Does it matter how this technology is implemented? Must students be engaged in a hands-on, active learning activity to benefit from this technology, or can the same understanding be attained by observing the instructor work through the components of the activity? What about universities without the advantage of a daily/weekly computer lab? Can these students benefit from observing the instructor work through the activity? This paper will highlight key findings from a classroom research pilot study carried out in Spring 2011 using a crossover design. This study investigates whether active participation in computer lab activities enhances student understanding of key statistical concepts, such as p-values and statistical significance, as compared to observing an instructor work through those same components of the activity. Results of assessment, along with student attitudes and opinions will be shared.

Exploring The Role Of Problem Context In Students' Understanding Of Sampling

Herle McGowan, North Carolina State University; ✦Leigh Slauson, Capital University, , *lslauson@capital.edu*; Tara Cope , SUNY-Adirondack; Jacqueline Wroughton, Northern Kentucky University

Key Words: context, sampling, inference

Context plays an important role in the discipline of Statistics as providing meaning for data analysis and the evaluation of evidence, but may be distracting to students. The use of good sampling methods also plays an important role in Statistics, as necessary for ensuring appropriate inference. This research explores the role of context in students' reasoning about sampling. Data was collected at four diverse academic institutions and analyzed to explore the interaction between the strength of a student's opinion about a topic--providing the context for a study--and their ability to judge the quality of the sampling method and the scope of the conclusions in the study. Results will be presented and implications for future research discussed.

A Comparison Of Students' Inferential Reasoning In Three College Courses

✦Sharon Lane-Getaz, Saint Olaf College, 1520 Saint Olaf Avenue, Northfield, MN 55057 USA, *lanegeta@stolaf.edu*

Key Words: Statistics education, inference, assessment

This quasi-experimental study reports a comparison of inferential reasoning in three college statistics courses in spring 2010. A 35-item version of Reasoning about P-values and Statistical Significance (RPASS) scale was administered as a Pretest and Posttest to assess and compare gains in students' inferential reasoning. Gains were adjusted for prior knowledge and mathematics ability. RPASS-8 is a research tool designed to assess effects of different teaching methods on students' inferential reasoning. Score reliability was estimated at .98 (Cronbach's coefficient alpha). As evidence of construct-related validity RPASS-8 scores were correlated with college entrance scores and student-reported GPAs. In addition to quantitative evidence, students' inferential reasoning was analyzed for twelve selected items. Future development and research are discussed.

Comparison of Learning Outcomes and Attitudes for Students Completing Introductory Statistics Homework Online vs. Hardcopy vs. Self-Selecting from the Two Methods

✦Nancy Pfenning, University of Pittsburgh, 2316 Shady Ave., Pittsburgh, PA 15217, *nancypfenning@gmail.com*

Key Words: online, homework, self-selection, learning, attitudes, hardcopy

More and more instructors are taking advantage of online assignment completion for their students. An obvious gain for instructors is that the amount of time spent grading papers is drastically reduced. What benefits or drawbacks are there for students? If students were given a choice, would they prefer to complete their homework online or with the traditional hardcopy method? A carefully controlled comparative study by the author explores students' attitudes toward the two methods. In addition to a hardcopy-method lecture class and an online-method lecture class, a third lecture class will give students an opportunity to choose their preferred method. Because these are three back-to-back lectures of the same course, subtle differences in learning outcomes may be detected. Even if there is no significant difference in achievement for the three groups, there may well be important differences in students' level of satisfaction with their assigned or self-selected method. The results of this study may guide other instructors in their decision of whether to use online or hardcopy homework for their students, or perhaps to let the students themselves make the choice.

Assessing The Change In Student Attitudes Towards Statistics At The University Of Tennessee

✦Ramon Leon, University of Tennessee, SMC 337, 916 Volunteer Blvd, Knoxville, TN 37996 United States, *rleon@utk.edu*; Adam Petrie, University of Tennessee

Key Words: attitudes, introductory, classes, SATS-36, blog, tutors

The SATS-36© survey measures six aspects of student attitudes towards statistics and collects additional information about students' backgrounds. Pre-class and post-class scores can be analyzed to assess changes in attitudes. In this study, we examine these changes for approximately 950 students at the University of Tennessee enrolled in the Spring 2011 introductory statistics course, with about 900 from ten regular sections of the course and about 50 from two honors sections. Further, we examine the effect of intelligent tutors, online coaching sessions, a blog, and other tools in two of the regular and one honors sections to see if new methods make an impact in the change of attitude. Results are pending completion of the semester.

Conceptualizing And Measuring "Data Habit-Of-Mind"

✦Saad Chahine, OISE/university of Toronto, Toronto, ON L4S 2C1 Canada, *saad@metrein.com*

Key Words: Data Habit-of-Mind, Data-Based Report Interpretation, Statistical Literacy, Statistical Education, performance, education

Professionals are increasingly being asked to interact with data to enhance their performance in the workplace. However, there is little research that examines the factors that facilitate data use. This study proposed a model, Data Habit-of-Mind (DHoM), to represent professionals' use of data from large-scale assessments. The metaphor, Habit-of-Mind, was originally coined by Keating (1990) and represents individual's habits of inquiry. Based on an extensive review of the literature, DHoM was defined as the combination of two traits: Statistical Literacy and Data-Based Report Interpretation. To test this model, 20 educators were interviewed using the cognitive interview method (Willis, 2006). In the interview, educators engaged in performance-based tasks that assessed their level of Statistical Literacy and Data-Based Report Interpretation. Results were analyzed using a qualitative matrix method to group educators into different categories based on their performance on tasks. Findings showed the there were five types of educators based on the discrepancy in their scores. Implications are discussed in relation to professional development needs in the workplace.

The Effect of Student-Driven Projects on the Development of Statistical Reasoning

✦Melissa M. Sovak, California University of Pennsylvania, 317 Eberly College, California University of Pennsylvania, California, PA 15419, *Sovak@calu.edu*

Key Words: Statistics Education, Projects, Teaching

Research has shown that passing an introductory statistics course does not imply that a student can reason statistically. Since there is minimal experimental evidence in this area, the purpose of this study was to produce quantitative data from a designed comparative study to explore the effectiveness of a semester-long project in enhancing students' statistical reasoning abilities along with providing a template for these projects. The project was designed to target the areas of distributions, probability and inference. The study included two experimental sections and five control sections. All sections completed both a pre-test and post-test designed to measure reasoning ability. Analysis of the data was completed using ANCOVA, contingency tables and a qualitative analysis to investigate the effect of the project on the development of statistical reasoning. Analysis of the data indicated that project participants had higher learning gains overall and a more enjoyable experience in their statistics course when compared with students not participating in the project. These results indicate that projects are a valuable teaching technique for introductory statistics courses.

420 Modeling Pollution and Air Quality

■●

Section on Statistics and the Environment, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Calibration Of Air Quality Deterministic Models Using Nonparametric Spatial Density Functions

✦Jingwen Zhou, Dept of Statistics, North Carolina State University, 2804 Brigadoon Dr. APT 22, Raleigh, NC 27606, *jzhou3@ncsu.edu*; Montse Fuentes, North Carolina State University

Key Words: Model calibration, Quantile regression, Nonlinear monotonic regression, Nonparametric Bayes method

The evaluation of physically based computer models for air quality applications is crucial to assist in control strategy selection.The objective comparison of mean and variances of modeled air pollution concentrations with the ones obtained from observed field data is the common approach for assessment of model performance.One drawback of this strategy is that it fails to calibrate properly the tails of the modeled air pollution distribution,and improving the ability of these numerical models to characterize high pollution events is of critical interest for air quality management.In this work we introduce an innovative framework to assess model performance,not only based on the two first moments,but on their entire distribution.Our approach also compares the spatial dependence and variability in both models and data. More specifically,we estimate the spatial quantile functions for both models and data,and we apply a nonlinear monotonic regression approach on the quantile functions taking into account the spatial dependence to compare the density functions of numerical models and field data. We use a Bayesian approach for estimation and fitting to characterize uncertainties.

Power Calculation And Simulation For Studies Of Estimating Chronic Health Effect Of Pm2.5 Using Three-Level Hierarchical Random Effect Model

✦Ayano Takeuchi, Department of Biostatistics The University of Tokyo, International , *takeuchi@epistat.m.u-tokyo.ac.jp*

Key Words: power calculation, PM2.5, chronic effect, cohort study, three-level hierarchical model

Last decade, it has been known that fine particulate matter (PM2.5) has some acute health effect from the result of various time series studies. In this time, we have decided to evaluate chronic influence of PM2.5 on the development of child pulmonary function. We plan to conduct cohort study in more than 10 regions of Japan, and measure lung function and any other covariates for elementary school student for four years. We plan to use three-level hierarchical random effect model for adjusting individual development of lung function, time-related covariates, subject-specific covariate, and community-specific covariate to estimate the chronic effect of PM2.5. It is necessary to calculate power and sample size, especially the balance of the number of community and the number of subjects within one community. I did power calculation and simulation study for these purpose. In conclusion, when I fixed the number of subjects, the power increases according as the number of regions increasing, and we have enough power to follow up more than 10 regions and total 1000 subjects.

Multiplicative Factor Analysis With Latent Mixed Model For Exposure Assessment Of Airborne Particulate Matter

✦Margaret Claire Nikolov, United States Naval Academy, Department of Mathematics, 572C Holloway Road, Chauvenet Hall, Annapolis, MD 21402-5002, *nikolov@usna.edu*; Brent Coull, Biostatistics Department, HSPH

Key Words: source apportionment, latent variable, factor analysis, multiplicative error, mixed model

A major goal of air pollution research is to relate airborne particulate matter (PM) to speci?c sources which are often unmeasurable. Source apportionment and multivariate receptor modeling use standard factor analytic techniques to estimate source-speci?c contributions from a set of measured chemical species. We propose a multiplicative factor analysis with a mixed model structure on the latent source contributions. A factor analysis with multiplicative errors maintains the non-negativity of measured chemical concentrations, while the mixed model provides for systematic and random e?ects on the unobserved sources. We show that when applied to samples of ambient Boston aerosol, the multiplicative model provides better model fit over the standard additive model. Using this framework, we explore the effects of meteorological factors, such as wind direction and wind speed, on source-specific PM. Preliminary analysis of the Boston data indicates increased power plant PM associated with wind trajectories from the west/southwest, increased oil combustion PM associated with wind trajectories from the northwest, and elevated motor vehicle particles during stagnant air mass.

A Statistical Approach For Aerosol Retrieval Using Misr Data

✦Yueqing Wang, University of California at Berkeley, 2 Panoramic Way Apt 105, Berkeley, CA 94704 United States, *yqwang@stat.berkeley.edu*; Xin Jiang, Peking University; Ming Jiang, Peking University; Bin Yu, UC Berkeley

Key Words: Aerosol retrieval, Multi-angle Imaging SpectroRadiometer, Hierarchical Bayesian mode, Markov Chain Monte Carlo, Maximum a Posteriori, Remote sensing

Atmospheric aerosols when inhaled can penetrate cell membranes and cause serious damage to the cardiovascular system. In addition, black carbon aerosols produced by human activities can reduce ground-level visibility. Therefore, an accurate profile of aerosols' spatial distribution is very important for air pollution monitoring, especially for the highly-populated urban areas. NASA's Multi-angle Imaging SpectroRadiometer(MISR) has demonstrated the capability to retrieve global Aerosol Optical Depth(AOD) at a resolution of 17.6 km. However, to provide informative proxies for local air quality studies, a finer resolution is desirable. Building a hierarchical Bayesian model based on MISR data, we improve the current retrieval algorithm in three aspects. First, we retrieve AOD at a resolution of 4.4 km, with model validation by ground-based measurements in the Beijing area, China. Second, in modeling aerosol distribution over a geographical lattice, we incorporate a spatial dependence structure, motivated by the air particle interactions. Finally, we expand the set of possible aerosol compositions. Several case studies are conducted to discuss the importance of the above developments.

A Bayesian Spatio-Temporal Model For Estimating Daily Nitrogen Dioxide Levels

✦Lixun Zhan, Yale School of Public Health, , *lxzhang@gmail.com*; Yongtao Guan, Yale School of Public Health; Brian P Leaderer, Yale University; Theodore R Holford, Yale University

Key Words: Bayesian hierarchical model, longitudinal model, spatial statistics, daily nitrogen dioxide, Environmental Protection Agency

A Bayesian spatio-temporal model for estimating \emph{daily} nitrogen dioxide (NO_2) levels is described. The model uses two datasets with different temporal resolutions. The dataset from the Study of Traffic, Air quality and Respiratory health in children (STAR) contains NO_2 measurements at a relatively large number of sites (most of which are in the state of Connecticut) but for just a few \emph{monthly} observations at each site. The dataset from the U.S. Environmental Protection Agency (EPA) contains measurements on an \emph{hourly} level but only at a limited number of sites (four sites in Connecticut). The model first establishes relationship between STAR observations and EPA observations on the monthly level (the duration-level of the STAR study). The relationship is then assumed to hold at the daily level and thus daily NO_2 levels at the STAR study sites could be estimated from the average daily NO_2 levels at the EPA sites. The model can also provide predictions of daily NO_2 levels at random sites. The model is implemented under the Bayes framework with a Gibbs sampler and performed well ($R^2>0.7$).

Dynamic Spatio-Temporal Models For The Diurnal Cycle

✦Jonathan Hobbs, Iowa State University, 2804 Stange Rd #8, Ames, IA 50010, *jonhobbs@iastate.edu*

Key Words: dynamic model, spatio-temporal, spatial statistics, meteorology

Many environmental processes exhibit a characteristic diurnal cycle, which can vary in complex ways in space and time. Spatio-temporal models with dynamically-evolving parameters are flexible tools for characterizing the evolution of the diurnal cycle. Bayesian analysis of models of varying complexity reveals combinations of parameters that exhibit a diurnal cycle and that vary spatially or from day-to-day. Applications to regional fields of atmospheric humidity and precipitation are presented. Day-to-day variability in the large-scale mean and small-scale variability are noticeable in these moisture fields for the region of interest.

421 Statistical Methodology with Complex Correlation Structures With Applications in Epidemiology

Section on Statistics in Epidemiology, ENAR, Section on Health Policy Statistics, Scientific and Public Affairs Advisory Committee

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Network Models For Studying Frailty As A Dynamic System

✦Hong Zhu, Division of Biostatistics, Ohio State University College of Public Health, 320 W. 10th Ave., B-118 Starling Loving Hall, Columbus, OH 43210 USA, *hzhu@cph.osu.edu*; Ravi Varadhan, Division of Geriatrics & Gerontology, Johns Hopkins School of Medicine

Key Words: network model, frailty, dynamic system, stimulus-response experiment

Frailty is a state of health signified by an increased vulnerability to adverse health outcomes in the face of stressors. There has been emerging interest on the research of frailty, which focuses on the dynamic interactions within and across the complex adaptive systems underlying the frailty syndrome. The theoretical literature on frailty hypothesizes that frail older adults are characterized by a loss of resilience in the face of stressors, and that the frail and non-frail would differ in terms of the dynamics of physiological systems in response to stimuli. It is challenging to formalize the notion of resilience. It is a characteristic of the dynamics displayed by the system in response to a variety of stimuli. We propose network models for studying frailty as a dynamic system, based on the stimulus-response experiment. We demonstrate that resilience of a complex system can be quantified and modeled in different ways. Through simulation studies, we explore the relationships among network structure, its topological properties, and the dynamics of stimulus response.Our work should provide a deeper understanding of the role of multisystem interactions in the etiology of frailty.

Testing On The Multivariate Normal Covariance Matrix In High-Dimensions

✦Thomas Fisher, University of Missouri-Kansas City, Department of Mathematics & Statistics, 5100 Rockhill Road, Kansas City, MO 64110-2499, *fishertho@umkc.edu*

Key Words: Hypothesis Testing, Covariance matrix, High-dimensional data analysis

In this presentation we will discuss the problem of hypothesis testing on the covariance matrix when the dimensionality exceeds the sample size; a problem motivated by DNA microarray data. When in high-dimensions, the sample covariance matrix is singular and the likelihood ratio criterion is degenerate. Results in the literature recommend tests based on the arithmetic means of the eigenvalues of the covariance matrix as they are not perturbed by zeros. Much of the results rest on estimators for the first and second arithmetic means. We have developed unbiased and consistent estimators for the third and fourth arithmetic means of the eigenvalues. New test statistics are proposed under comparable assumptions to those statistics in the literature for the identity and sphericity hypotheses. The asymptotic distribution of the proposed test statistics are found in the general asymptotic framework. The statistics are shown to be consistent and comparable to those in the literature. A simulation study shows the newly proposed tests are effective and more powerful than those in the literature when just a few elements deviate from the null hypothesis.

Modeling Changes In Bacterial Prevalence With Sequenced Data

✦Raymond G Hoffmann, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI 53226, *rhoffmann@mcw.edu*; Ke Yan, Medical College of Wisconsin; Jessica VandeWalle, University of Wisconsin Milwaukee; Sandra McLellan, University of Wisconsin Milwaukee

Key Words: wavelet analysis, harmonic regression, time series, normalization

Information about the prevalence of bacterial taxa in water samples can be determined via next generation sequencing. Samples were taken from two different, but related, sites in lake Michigan over a three year period. Since the samples were taken during different seasons and at different points in time, determining temporal effects required methods that were both flexible and did not required regular measurements. Initially there were over 600 taxa; after aggregation of similar taxa there were still 22 time series with day specific and site specific covariate information. Wavelet analysis after filtering to reduce sampling noise was used to determine the temporal aspects of the multivariate time series because of the irregular occurrence of the sampling points over time. Initial attempts with sine-cosine time series failed because of the irregularity of the time points and shapes of the prevalence curves. Since the data was obtained as separate samples, different methods of normalization were also explored. The main goal was to determine if wavelet coefficients could be used to identify common patterns of temporal response among the major taxa and sub-taxa.

Treatment - Outcome Complex And Analysis Of Observational Data

✦Lev S Sverdlov, Independent Researcher, 07081, *lev.sverdlov@gmail.com*

Key Words: Treatment - Outcome Complex, Observational Data, Longitudinal, Micro-experiment, Intended and Unintended Effects, Inferences

The goal is to create a logical framework for analysis of intended and unintended effects of treatment using a concept of the treatment-outcome complex (TOC) (Sverdlov, 1996, 2004) for analysis of longitudinal observational data. This approach assumes that at any moment during treatment of chronic conditions we in fact can assess neither a severity of the disorder, nor an outcome of treatment. Rather it is a momentary functional state of the patient under the influence of therapy and under a condition described by a set of observed and unobserved covariates as they stand at this moment (TOC). The process of treatment then can be described as a sequence of TOCs. The change in the outcome component of each sequential TOC depends on the change in treatment, i.e. each sequential pair of adjacent TOCs within an individual patient record can be interpreted as a micro-experiment. Various temporal patterns of outcomes can be analyzed using relevant aggregations of the TOCs with a possibility of measuring the effects of history and maturation. The approach can be used for assessing treatment effects in changing environments.The opportunities and limitations of the approach will be discussed

Analysis Of Zero-Inflated Count Time Series: A Partial Likelihood Approach

✦Ming Yang, University of Iowa, , *ming-yang@uiowa.edu*; Gideon Zamba, Department of Biostatistics; Joseph Cavanaugh, University of Iowa

Key Words: Count time series, EM algorithm, Partial likelihood, Zero-inflation

Count data with excess zeros are common in many biomedical and public health applications. The zero-inflated Poisson (ZIP) regression model has been widely used in practice to analyze such data. In this paper, we extend the ordinary ZIP regression framework to model zero-inflated count time series. An observation-driven model is presented and developed, and the partial likelihood is employed for statistical inference. Partial likelihood inference has been successfully applied in modeling time series where the conditional distribution of the response lies within the exponential family. Extending this approach to ZIP time series poses methodological and theoretical challenges, since the ZIP distribution is a mixture and therefore lies outside the exponential family. We establish the asymptotic theory of the maximum partial likelihood estimator (MPLE) under mild regularity conditions, and investigate its finite sample behavior in a simulation study. We outline the computation of the MPLE and its standard error. Finally, we present an epidemiological application to illustrate the proposed methodology.

Current Status Observation Of A Counting Process With Application To Simultaneous Accurate And Diluted Assay Hiv Test Data

✦Karen McKeown, University of California, Berkeley, 101 Haviland Hall, University of California, Berkeley, Berkeley, CA 94720, *karen.mckeown@berkeley.edu*; Nicholas P Jewell, University of California, Berkeley

Key Words: Current status data, multistate models, HIV test data

We examine multistate current status data defined by two survival times of interest where one only observes whether or not each of the individual survival times exceed a common observed monitoring time. An individual then belongs to one of three states. We are interested in whether current status information on the second event can be used to improve estimation of the distribution function of time to the first event. For both single and multiple monitoring time scenarios, in the fully nonparametric setting, one cannot improve the naive estimator, using information on the first event only, when estimating smooth functionals of the distribution of time to the first event (van der Laan & Jewell 2003). We examine improving this naive estimator when parametric assumptions about the waiting time between the two events are made. For situations where this waiting time is modifiable by design, the issue of determining the optimal length of the waiting time for estimation of the cumulative hazard of the distribution of time to the first event in the recent past is also addressed. The ideas are motivated by and applied to an example on simultaneous accurate and diluted assay HIV test data.

422 Novel Methodological Issues in Reproductive Epidemiology

Section on Statistics in Epidemiology, ENAR

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Infant Weight-Specific Mortality: Degenerate Mixture vs. Non-Degenerate Mixture

✦Eric Yaw Frimpong, FDA, RM 4651 WO 21, 10993 New Hampshire ave, Silver Spring, MD 20993, *eric.frimpong@fda.hhs.gov*; Timothy Gage, University at Albany, SUNY; Howard Stratton, University at Albany, SUNY

Key Words: Birth weight, Infant mortality, degenerate mixtures, non-degenerate mixtures

To estimate the level of unobserved heterogeneity in infant weight specific mortality, several mixture models have been developed. Birth weight has been established as the most predictive factor in infant mortality. The marginal of the covariate, birth weight has been considered in both a degenerate mixture and a non-degenerate mixture. There has been considerable interest in the role of birth weight and/or gestational period in the mixture formulation. To explore this question, real Birth/Death data sets obtained from National Center for Health Statistics (NCHS) and simulations are used to investigate total mortality based on the mixture formulations. Hazard rate functions such as Weibull, Gompertz and Gamma are considered. Mixture formulations of these different distributions are also examined by simulations.

Urban Land-Use and Infants' Respiratory Symptoms

✦Keita Ebisu, Yale University, LEPH 60 College Street #201, New Haven, CT 06520 US, *keita.ebisu@yale.edu*; Theodore R Holford, Yale University; Kathleen D Belanger, Yale University; Brian P Leaderer, Yale University; Michelle L Bell, Yale University

Key Words: Wheeze Symptom, Land-use, Infants' health, Traffic, Urbanicity

The relationship between urban land-use and health is not fully understood. We investigated whether urbanicity near residence is associated with infant's respiratory symptoms. Wheeze occurrence was recorded for first year of life for 680 infants in Connecticut ,1996-1998. The fraction of urban land-use near a subject's home, assessed through satellite imagery, was related to severity of wheeze symptoms using ordered logistic regression, adjusting for individual variables (e.g., race). NO2 exposure, as a proxy for traffic pollutants, was estimated using integrated traffic exposure modeling. A 10% increase in urban land-use within a 1,540m buffer of each infant's residence was associated with a 1.09-fold increased risk of severe wheeze (95% C.I., 1.02-1.16). When NO2, representing traffic pollution, was added to the model, results for urban land-use are no longer statistically significant, but have a similar central estimate. Findings indicate that urban land-use is associated with respiratory symptoms in infants, and that health effect estimates for urbanicity incorporate some effect of traffic-related emissions, but also involve other factors.

Bayesian Mixed Hidden Markov Models: A Multi-Level Approach To Modeling Outcomes With Misclassification

✦Yue Zhang, University of Southern California, , *zhangyue@usc.edu*; Kiros Berhane, University of Southern California

Key Words: Mixed Hidden Markov Model (MHMM), Multi-Level Model, Misclassification, MCMC, Bayesian, Asthma

Questionnaire-based health status responses are often prone to misclassification. When studying the effect of risk factors on such responses, ignoring the possible misclassifications may lead to biased effect estimates. Analytical challenges posed by these misclassified responses are further complicated when simultaneously exploring the factors for both misclassification and health process in a multi-level setting. We propose a fully Bayesian Mixed Hidden Markov Model (BMHMM) for handling differential misclassification in discrete responses in a multi-level setting. The BMHMM generalizes the Hidden Markov Model (HMM) by introducing random effects into three sets of HMM parameters for prevalence, transition and emission probabilities, allowing for cluster level heterogeneity under a multi-level model structure. An extensive simulation study is undertaken to illustrate the gains from properly accounting for the misclassification. We apply our method to the Southern California Children's Health Study, where questionnaire based information on asthma diagnosis in children may be observed with misclassification. Risk factors for both asthma transition and misclassification are examined.

Structured Additive Regression (Star) Modelling Of Women'S Age Of Menarche And Years Of Fertility In Central Portugal

✦Bruno C de Sousa, Institute of Hygiene and Tropical Medicine, Rua da Junqueira 100, Lisboa, 1349-008 Portugal, *bruno.desousa@ihmt.unl.pt*; Elisa Duarte, University of Santiago de Compostela; Carmen Cadarso-Su·rez, University of Santiago de Compostela; VÌtor Rodrigues, University of Coimbra; Thomas Kneib, Institut f¸r Mathematik, Carl von Ossietzky University Oldenburg

Key Words: Structured Additive Regression (STAR), Breast Cancer, Screening program, Biostatistics

The Portuguese Cancer League (LPCC) is a private non-profit organization dealing with multiple issues related to oncology, including the National Breast Cancer Screening Program. In this study, we analyze approximately 260,000 data records of women that entered the Screening Program for the first time in the central region of Portugal. It is believed that the period of time between the age of menarche and the age of menopause has been increasing over time. Therefore, a new variable called Window was defined as the difference between the age of menopause and the age of menarche, which represents a woman's years of fertility. In this study, the evolution in time and space of the variables Window and the Age of Menarche were analyzed through STAR models, exploring the possible associations with other variables, such as Hormone Replacement Therapy, Pregnancy Status, Nursing Status, and Contraceptive Pills.

Bayesian Order Restricted Inference For Hormonal Dynamics

✦Michelle Renee Danaher, Eunice Kennedy Shriver, National Institute of Child Health and Human Development Epidemiology Branch, 6100 Executive Blvd, Rm 7B03, Rockville, MD 20854 USA, *danahermr@mail.nih.gov*; Anindya Roy, University of Maryland Baltimore County; Sunni L. Mumford L. Mumford, Eunice Kennedy Shriver, National Institute of Child Health and Human Development; Enrique F. Schisterman, Eunice Kennedy Shriver, National Institute of Child Health and Human Development; Paul S Albert, Eunice Kennedy Shriver, National Institute of Child Health and Human Development; Zhen Chen, Eunice Kennedy Shriver, National Institute of Child Health and Human Development

Key Words: Bayesian procedures, Hormone measurements, Menstrual cycle, Order-restricted inference, Shape constraints

Biomedical data often arise from well-understood biological processes. For example, hormone levels during the menstrual cycle are driven by well-established biological feedback mechanisms between luteinizing hormone, follicle-stimulating hormone, progesterone, and estrogen. Incorporating the known restrictions imposed by the underlying biological processes into the statistical model can greatly improve statistical efficiency and provide estimates of factors affecting menstrual cycle function that are interpretable within the context of known biological relationships. To address these constraints we propose a Bayesian procedure by specifying priors on the constraint space using a reparameterization via Minkowski decomposition. We perform simulations to investigate properties of the proposed methods and for comparison use an existing procedure that is similar to a Bayesian procedure, in which draws from an unconstrained posterior distribution are projected to a constraint space via optimization methods. Lastly, we demonstrate application of these methods to hormone data from the BioCycle study.

Alcohol Consumption During Pregnancy And Risk Of Placental Abruption And Placenta Previa.

✦ONeil Lynch, Minnesota State University Moorhead, Mathematics Department, Minnesota State University Moorhead, Moorhead, MN 56563 USA, *lynch@mnstate.edu*

Key Words: Alcohol, Placenta previa, Placental abruption, Population-based study

The main outcomes of interest were placenta previa, placental abruption and a composite outcome defined as the occurrence of either or both lesions. Multivariate logistic regression was used to generate adjusted odd ratios, with non-drinking mothers as the referent category. Women who consumed alcohol during pregnancy had a 33% greater likelihood for placental abruption during pregnancy (adjusted odds ratio (OR), 95% confidence interval (CI) = 1.33 [1.16-1.54]). No association was observed between prenatal alcohol use and the risk of placenta previa. Alcohol consumption in pregnancy was positively related to the occurrence of either or both placental conditions (adjusted OR [95% CI] = 1.29 [1.14-1.45]). Mothers who consumed alcohol during pregnancy were at elevated risk of experiencing placental abruption, but not placenta previa. Our findings underscore the need for screening and behavioral counseling interventions to combat alcohol use by pregnant women and women of childbearing age.

Adjustment For Bias Using Propensity Scores: A Simulation Study Of Receipt Of Respiratory Syncytial Virus (Rsv) Immunoprophylaxis To Prevent Rsv Infection And Infant Outcomes

✦Pingsheng Wu, Vanderbilt University School of Medicine, 1161 21st Ave. South, S2323 MCN, Nashville, TN 37232, *wupingsheng@hotmail.com*; William D Dupont, Vanderbilt University School of Medicine; Gabriel Escobar, Kaiser Permanente Northern California ; Tina Hartert, Vanderbilt University School of Medicine; Eileen M Walsh, Kaiser Permanente Northern California ; Kecia Carroll, Vanderbilt University School of Medicine; Sherian Li, Kaiser Permanente Northern California; Edwards Mitchel, Vanderbilt University School of Medicine; Tebeb Gebretsadik, Vanderbilt University School of Medicine ; Jeff Horner, Vanderbilt University School of Medicine ; Patrick Arbogast, Vanderbilt University School of Medicine

Key Words: Propensity score, Bias adjustment

Concern exists that physicians may prescribe respiratory syncytial virus (RSV) immunoprophylaxis to only selected infants within the established high-risk groups, resulting in bias in estimating the effect of RSV immunoprophylaxis on RSV-attributable morbidity. We conducted a retrospective cohort study of 1205 privately insured premature (<32 weeks) infants who were eligible for and in whom RSV immunoprophylaxis is recommended by the American Academy of Pediatrics. Seventy seven percent of infants who were eligible received at least one dose of RSV immunoprophylaxis, and 5% experienced a RSV related hospitalization. Simulation studies were conducted to compare the performance of four propensity score methods in reducing bias in receipt of RSV immunoprophylaxis: stratification on the propensity score, propensity score matching, inverse probability of treatment weighting using the propensity score, and covariate adjustment using the propensity score. Performance of the methods relative to the true underlying association was compared and reported for the motivated RSV immunoprophylaxis and morbidity research.

423 Contributed Oral Poster Presentations: Biopharmaceutical Section

Biopharmaceutical Section

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

A Joint Model for Tumor Burden and Progression-Free Survival

✦Aparna B. Anderson, Bristol-Myers Squibb Company, Global Biometric Sciences, 5 Research Parkway, Wallingford, CT 06492, *aparna.anderson@bms.com*; Ye Shen, Yale University School of Public Health, Division of Biostatistics; Ritwik Sinha, Hewlett-Packard India, Decision Support and Analytics Services

Key Words: longitudinal data, progression-free survival, mixed effects, oncology

In oncology, overall survival is the ideal measure of treatment benefit. However, the mechanistic and biologic effects of a therapeutic agent are generally characterized as changes in tumor burden (TB) measured repeatedly at specified time points. In clinical trials that support marketing approval, TB is usually analyzed as a categorical variable (e.g., applying RECIST criteria) to summarize objective response rate or progression-free survival (PFS). To address the loss of information due to categorization as well as potentially informative missing TB, a joint model (Henderson, et al., 2000) for the longitudinal tumor burden process and PFS is considered. The operating characteristics of this model are evaluated under different tumor burden distributions and missing data mechanisms. Simulations indicate that joint modeling reduces estimation bias when data are missing not at random and leads to minimal loss of efficiency when conditions are consistent with linear mixed effects modeling assumptions.

Predict The Probability Of Final Trial Success In Interim Analysis

✦Rujun Teng, Merck & Co., Inc., , *rujun_teng@merck.com*; Kaifeng Lu, Forest Laboratories Inc

Key Words: futility, posterior probability, predictive probability

In clinical trials, enrolment of patients is a continual process staggered in time. In Phase II POC (proof of concept) trials, interim analyses are often performed to assess the available data, to find whether there is significant treatment difference, and to evaluate whether the difference is convincing enough to draw a conclusion. Since clinical trials are expensive, if superiority or futility conclusion can be reached at interim look earlier than completion of the whole trials, the cost will be less than planned. Based on the usual Bayesian framework, a posterior probability distribution for some parameter of interest (e.g. treatment effect) can be derived from the observed data at the interim analysis and a prior probability distribution for the parameter. For example, the conclusion treatment A is superior to treatment B is reached if the posterior probability is larger than some prespecified cutoff value. Alternatively, predictive probability can be used to calculate the conditional probability of success of final trial, given the available interim data. Simulation data will be used to compare the posterior probability and predictive probability approaches.

Using Z-Prime Factor As Quantitative Measure Of In Vivo Pd Assay Quality

✦Winnie Weng, AMGEN, 1201 Amgen CT W, MS/D3262, Seattle, WA 98119, *nweng@amgen.com*; Guang Chen, AMGEN; Margaret Weidner, AMGEN; Kathy Keegan, AMGEN; Christophe Queva, AMGEN

Key Words: z prime factor, PD assay, pharmacodynamic assay, in vivo

In vivo pharmacodynamic (PD) assays have been used extensively in preclinical studies of drug development. They are used to develop a correlation between the mechanism of action of a drug with its desired efficacious outcome. To date, in vivo screening assays are evaluated independently without a standard method to compare one to another. Here we present the development of Z prime factor; a simple statistical parameter reflecting the combined effect of width of the assay window (i.e. difference between positive and negative controls), variation associated with assay, and number of animal and/or technical replicates used. Using Z prime factor as a quantitative measure of in vivo PD Assay quality, investigators can determine if the assay is sensitive enough for drug compound screening, as well as to estimate the fewest number of animals needed to achieve statistically significance, and to compare quality of different assay platforms.

Analysis Of Gef-Mediated Nucleotide Exchange Assay Experiments

✦Qinghua Song, Genentech, 94080, *song.qinghua@gene.com*; Lindsay Garrenton, Genentech; Imola Fodor, Genentech; Guowei Fang, Genentech; Peter Jackson, Genentech

Key Words: GEF, exponential decay model, four-parameter logistic fit

Guanine nucleotide exchange factors (GEF) are components of intracellular signaling networks that activate downstream proteins by stimulating the exchange of guanosine diphosphate (GDP) for guanosine triphosphate (GTP). Son of Sevenless (SOS) is a Ras-specific GEF that activates Ras proteins, which play a role in regulating cell growth, differentiation, and survival, and have been implicated in many types of cancer. To biochemically characterize GEF-mediated nucleotide exchange, we developed fluorescence-based GEF assays and statistical methods for their analysis. We fit an exponential decay model to the time series of each enzyme concentration and estimate the corresponding exponential decay rates. We then display the ratio of estimated rates versus that of no enzyme control as a function of increasing log enzyme concentration and estimate the EC50 for the exchange reaction via a four-parameter logistic fit. We illustrate the data analysis steps and results with an assay for a titration of the catalytic subunit of the Ras-GEF, SOS1. The automated data analysis process we developed will facilitate the interpretation of multiple future experiments.

Using Reinforcement Learning Strategies To Discover The Optimal Treatment For Advanced Colorectal Cancer Patients

✦Zheng Ren, The University of North Carolina at Chapel Hill, 1000 Smith Level Rd. Apt. B-9, Carrboro, NC 27510, *renzheng@email.unc.edu*; Michael R Kosorok, UNC-CH

Key Words: Personalized Medicine, Biomarkers, Colorectal cancer, Clinical trials, Dynamic treatment regime, Reinforcement learning

Reinforcement learning methods have been developed to identify individual dynamic treatment regimens for cancer patients with the ability to identify the optimal treatment strategy from a complex clinical setting, including selecting from several first line treatment options, several second line treatment options and the time of initiating second line

treatment. In this study, we use reinforcement learning to analyze data from a colorectal cancer trial. Biomarkers are believed to be useful for improving treatment and prognosis, so the optimal dynamic treatment rule was determined for individuals based on their clinical factors and biomarkers using reinforcement learning. The best treatment plan obtained from reinforcement learning is then compared with results from other methods in a simulation study, demonstrating that reinforcement learning is useful for discovering personalized medicine using a good clinical trial design with a reasonable sample size.

A Distribution-Free Bayesian Method For Estimating The Probability Of Response In Combination Drug Tests

✦John W Seaman III, Alcon Laboratories, 76134, *john.seaman@alconlabs.com*; John W Seaman, Jr, Baylor University; James Stamey, Baylor University

Key Words: Early Phase Studies, Prior Construction, Proof Loading

Many illnesses are often treated with a combination of drugs. These combinations can be more effective than using any of the component drugs individually, but may lead to increased safety concerns. Prior to human trials with the combination, what can be said about efficacy and/or safety of the combination? An experimental design known as proof loading allows us to obtain preliminary estimates about the joint probability of an adverse event, without exposing patients to the combination drug. We propose a Bayesian distribution-free approach to proof loading as a possible solution to this problem. We investigate the model with examples and simulation studies.

An Application Of Group Sequential Method For Demonstration Of Efficacy In Diagnostic Imaging Development

✦Gajanan Bhat, Lantheus Medical Imaging, 331 Treble Cove Road, North Billerica, MA 01862, *Gajanan.Bhat@lantheus.com*; Jeffrey Joseph, Omnicare Clinical Research, Inc.; Dana Washburn, Lantheus Medical Imaging, Inc.

Key Words: Sensitivity, Specificity, diagnostic efficacy, interim analysis, adaptive design, imaging trial

A joint hypothesis of sensitivity and specificity in a pivotal new diagnostic imaging product development is used to demonstrate efficacy. A new diagnostic imaging product (PET myocardial perfusion) is compared to an FDA-approved comparator imaging modality (SPECT) as well as both modalities being compared to an acceptable truth standard (coronary angiography (CA)). Where clinically the sensitivity is more important than specificity, the test can be one of superiority with regards to sensitivity, but one of non-inferiority with regards to specificity. Adaptive design approach (maximize statistical power and an interim analysis) in order to provide flexibility to Phase 3 will be critical to the success of the clinical program. A GSM for a joint hypothesis of superiority in sensitivity and non-inferiority in specificity incorporating O'Brien-Fleming function will be used. The total power of the study is the combination of powers attributed to individual hypothesis with the two parameters being interrelated and dependent on disease prevalence in trial population as determined by the CA. The proposed methodology provides an approach to control overall Type I error in this study.

Study Design Of A Drug-Drug Interaction Trial Using Group Sequential Methods

✦Wei Zhang, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, *zhang_wei_wz@lilly.com*

Key Words: group sequential methods, drug-drug interaction, sample size calculation, interim analysis

Drugs can interact with each other in a number of ways. The most common interaction, metabolic drug interaction, relates to how the liver metabolizes the drug substances. The substantial changes in exposure arising from metabolic drug interaction can alter the safety and efficacy profile of a drug in important ways. If sponsor wish to make a specific claim in product label that no drug-drug interaction is expected, then a confirmatory trial will need to be conducted to draw the conclusion. In general, a drug-drug interaction study requires 20-30 subjects to complete. This sample size would be a moderate number for healthy volunteers, but it usually causes considerable enrollment problem when an oncology compound is studied on cancer patients, due to the low expected enrollment rate. I propose a study design using group sequential methods to evaluate drug-drug interaction. This design allows an interim analysis that could possibly stop the study early for conclusion. Study design considerations, methods, sample size calculation, planned analyses, and computing, will be discussed.

A Comparison Of Methods For Adjusting For The Baseline Measure

✦Martin Ove Carlsson, Pfizer Inc., 235 East 42nd St., New York, NY , *martin.carlsson@pfizer.com*; Kelly H. Zou, Pfizer Inc; Ching-Ray Yu, Pfizer Inc; Franklin W. Sun, Pfizer Inc

Key Words: Clinical Trial, Baseline Measure, Change Score, Percent Change from Baseline, Scale Invariant, Analysis of Covariance

When analyzing randomized controlled clinical trials from baseline to the end of the study, we may encounter that within-subject baseline measures can impact inferential results once they are adjusted, compared with results from unadjusted methods. Various statistical methods may be employed to adjust for the baseline measure. For example, post-treatment score or change from baseline may be analyzed with the baseline measure as a covariate, along with additional available baseline covariates. The analysis of covariance is a common regression method for this adjustment. Alternatively, the percent change from baseline, a dimensionless quantity that is scale invariant, can be used as a covariate. The percent change is a convenient measure for examining the entire distribution of the responses. However, its exact distribution is not trivial, particularly when baseline and post-treatment scores are correlated. We propose improved approximations of the distribution of percent change from baseline using monotone nonlinear transformations. Monte-Carlo simulations are conducted to investigate and compare these adjustment methods. These methods are illustrated on multi-center trails.

Direct Cost Of Schizophrenia In Quebec, Canada: An Incidence-Based Microsimulation Monte-Carlo Markov Model

✦Alice Dragomir, University of Montreal, QC H3C3J7 Canada, *elena.alice.dragomir@umontreal.ca*; Jean-Eric Tarride, McMaster University; Ridha Joober, McGill University; Guy Rouleau,

University of Montreal; Sylvie Perreault, University of Montreal

Key Words: costs of schizophrenia, Markov Model with Monte-Carlo microsimulations, healthcare cost, non-healthcare cost

Aim: Pharmacological strategies for schizophrenia have received increasing attention due to the development of new and costly drug therapies. Objectives: To estimate the direct healthcare and non-healthcare cost of schizophrenia and to simulate cost reductions potentially obtained with a new treatment, over the first 5 years following their diagnosis. Methods: A microsimulation Monte-Carlo Markov model was used. Six discrete disorder states defined the Markov model. Costs and individual probabilities of transition were estimated from the administrative databases and all analyses were performed under the government perspective. Results: A total of 14,320 individuals were identified in the study cohort as newly diagnosed patients with schizophrenia. The mean direct cost of schizophrenia over the first 5 years following diagnosis was estimated $36,701. In the case where a new treatment with 20% increase of effectiveness will be available, the direct healthcare and non-healthcare costs can be reduced up to 14.2%. Conclusion: This model is the first Canadian model incorporating transition probabilities adjusted for individual risk-factor profiles and costs using real-life data.

Evaluation Of Power Of Different Cox Proportional Hazards Models Incorporating Stratification Factors

✦Shanshan Ding, University of Minnesota, 313 Ford Hall 224 Church Street S.E., Minneapolis, MN 55455, *pkususan@gmail.com*; Moumita Sinha, Biocon Ltd

Key Words: Cox proportional harzards model, survival time, stratification, simulation

Cox proportional hazards model is widely used in clinical trials for time to event data. Often mega trials have patients randomized by multiple stratification factors. These stratification factors are incorporated in the statistical analysis by using stratified Cox model. In this paper, we compare the stratified Cox model with the Cox proportional model using the stratification factors as covariates to study how the different models influence the power. Simulated data with and without stratification factors are explored with different event rates and variable number of patients in each subcategory. We conclude that fitting a Cox model with the stratification factors as covariates provides higher power compared to modeling the data with stratified Cox proportional model, especially when data structures are more differentiable among strata. This result is also observed when there are subcategories with low sample size.

Statistical Identifiability And Its Switching Mcmc Method On Nonlinear Pharmacokinetics Models

Seongho Kim, University of Louisville; ✦Hyejeong Jang, University of Louisville, 485 E. Gray ST., Louisville, KY 40222, *h0jang01@louisville.edu*; Lang Li, Indiana University School of Medicine

Key Words: Convergence rate, Michaelis-Menten (MM) kinetics, Monte Carlo Markov Chain (MCMC), Statistical Identifiability, Pharmacokinetics (PK), Switching algorithm

We study the convergence rate of MCMC on the statistically unidentifiable nonlinear model involving the Micaelis-Menten kinetic equation. We have shown that under certain condition, the convergence diagnosis by Raftery and Lewis (1992) is consistent with the convergence rate argued by Brooks and Roberts (1999). Therefore, different MCMC schemes developed in linear models are further extended and compared to the nonlinear models. We demonstrate that a single component MCMC (SCM) scheme is faster than the group component MCMC (GCM) scheme on unidentifiable models, while GCM is faster than SCM when the model is statistically identifiable. A novel MCMC method is then developed using both SCM and GCM schemes, which is called the Switching MCMC (SWM) method. The proposed SWM possesses the advantage of not having to know the identifiability of a model and, as a result, of being able to estimate parameters regardless of the statistical identifiable situations. In addition, simulations and data analysis suggest a better performance of the proposed SWM algorithm than SCM and GCM.

Stratified Two-Sample Tests On The Change Scores When The Outcomes Are Correlated Between Baseline And Post-Treatment: A Simulation Study Using Bivariate Data

Ching-Ray Yu, Pfizer Inc; ✦Martin Ove Carlsson, Pfizer Inc., 235 East 42nd St., New York, NY , *martin.carlsson@pfizer.com*; Kelly H. Zou, Pfizer Inc

Key Words: change score, two-sample tests, van Elteren's test, receiver operating characteristic analysis, Monte-Carlo, bivariate data

Multi-center randomized placebo control trials typically involve measuring a continuous outcome at baseline and post-treatment in the presence of strata. When the baseline and post-treatment scores are correlated, the change score may be computed. With the presence of strata, adjustment methods using different weights across strata may be applied. We investigate and compare the performances of nonparametric and parametric two-sample tests using change from baseline between the active medication and placebo control groups, after adjustment for strata. Nonparametrically, the van Elteren test using ranks of the data are conducted, using either local best or design free weights. Parametrically by assuming a normal distribution of the change scores, the McClish inverse-variance weight is applied to compute the area under the curve based on receiver operating characteristic analysis. Monte-Carlo simulations are conducted to compare and examine these tests using different weighting schemes by simulated bivariate normally or non-normally distributed data between baseline and post-treatment. The effect due to the correlation between the measurements at the two time points is evaluated

Leveraging Baseline Information To Improve Inference In Adaptive Randomized Experiments With Small Sample Size

✦Po-Han Brian Chen, Department of Biostatistics, Johns Hopkins University, 615N Wolfe Street, Room E3148, Baltimore, MD 21205, *pochen@jhsph.edu*; Rosenblum Michael, Department of Biostatistics, Johns Hopkins University

Key Words: Covariate-adaptive design, Confidence interval, Baseline information, Small sample, Randomized trial

We investigate how baseline variables can be used to improve inference in randomized experiments with small sample size. We focus on constructing small sample 95% confidence intervals for the mean treatment effect. When there is no baseline information (or it is ignored), standard exact methods for constructing confidence intervals can be used but they are generally conservative. We aim to incorporate baseline information to construct valid confidence intervals with shorter widths. First, we explore how much efficiency can be gained using standard randomization but doing an analysis adjusting for the baseline variables. Second, we explore covariate-adaptive randomization that attempts to balance predictive variables between the study arms. In our algorithm for constructing confidence intervals, the more predictive the baseline variables are, the shorter the resulting confidence intervals. We do a simulation study to explore how much shorter we can make our confidence intervals than those from exact methods that ignore baseline variables, and still get correct coverage probability.

Evaluating Oncology Phase I Dose Finding Study Designs

Lindsey Lian, PPD; ✦Hui Liu, PPD, 78744 U.S., *hui.liu@ppdi.com*; Jenny Huang, PPD

Key Words: study design, oncology, dose finding

The 3 + 3 design is still commonly used in oncology Phase I dose finding studies bearing its well known criticisms. To overcome the criticisms, alternative design methods have been proposed, including but not limited to modified continual reassessment method (MCRM) and escalation with overdose control (EWOC). Depending on the objectives and limits of a specific trial, different design method may fit the needs better. By using simulated data that mimic results from actual Oncology Phase I studies, we evaluate each design (3+3, MCRM, EWOC, etc) against different clinical trial setting, e.g. when duration/sample size of a trial is set, when trial subject safety/efficacy is of the most concern, when selecting the right phase II/III dose is the main goal, etc. Suggestions on best design fitting each setting are provided.

The Vacs Risk Index Responds To Treatment Interventions And Is Highly Correlated With And Predictive Of Mortality Events In The Optima Study

✦Katherine Anne Kirkwood, VA Cooperative Studies Program, 950 Campbell Ave, Building 35, 151A, West Haven, CT 06516, *katherine.kirkwood@va.gov*; Tassos Kyriakides, VA Cooperative Studies Program; Sheldon T. Brown, James J Peters VAMC, Infectious Disease Section, Bronx; Amy C. Justice, VA Connecticut Healthcare System, West Haven; Mark Holodniy, VA Palo Alto Health Care System; Janet Tate, VA Connecticut Healthcare System, West Haven; Joseph Goulet, VA Connecticut Healthcare System, West Haven

Key Words: randomized trials, mortality, prognostic index, biomarkers, treatment

Background: Reliable intermediate outcome measures are needed to assess clinical interventions in the current era of HIV treatment. We evaluated the performance of the Veterans Aging Cohort Study (VACS) Index which uses biomarkers to predict mortality. Methods: VACS Index scores were determined from data collected in the Options In Management with Antiretrovirals (OPTIMA) multi-national study of treatment strategies in patients with advanced HIV. Mean scores by treatment arms were compared using repeated measures analysis. Logistic regression and survival analysis were carried out to compare index score levels. Proportional hazards regression is being explored to assess the utility of the index score as an outcome measure for future clinical trials. Results: Log-rank comparisons of score strata highly correlated with mortality ($p < .0001$). The index performed well when applied to the OPTIMA data ($c = .728$). Conclusions: The VACS Risk Index accurately predicted mortality and responded to changes in treatment. It may offer an efficient alternative endpoint for the design of randomized clinical interventions among patients with advanced HIV.

Subgroup Discovery Showing Maximum Difference Between Treatment And Control Using Modified Patient Rule Induction Method (M-Prim)

✦Daniel Parks, GSK, , *daniel_c_park@gsk.com*; Xiwu Lin, GSK; Kwan Lee, GSK

Key Words: PRIM, M-PRIM, Subgroup

Identifying patient subgroups that show discernable differences in treatment effectiveness and/or safety profiles based on demographic and baseline variables is of major interest to clinicians and patients. In this work, we have extended the capability of the Patient Rule Induction method (PRIM) to automatically handle the important application of searching for subgroups that have large difference in treatment effectiveness when two treatments are compared. Comparison to other related methods and the limitations of the proposed method will also be discussed. Simulations and analysis of an actual clinical trial data will be shown to illustrate the proposed method.

A Phase Ii/Iii Randomized Clinical Trial Design With Sequential Decision Rule Based On Multiple Primary Endpoints

✦Qian Shi, Mayo Clinic, , *shi.qian2@mayo.edu*; Daniel J Sargent, Mayo Clinic

Key Words: clinical trial design, multiple endpoints, sequential hypothesis testing

Early efficacy screening before definitive phase III study is necessary to prevent sub-optimal regimens being tested further. A seamless phase II/III study design provides the opportunity to improve efficiency and provide early efficacy screening. In many practical scenarios, co-primary endpoints are needed to be considered to capture the unique nature of a disease. We proposed and applied a phase II/III randomized design with sequential final decision rule to a clinical trial comparing combined modality neo-adjuvant therapy to the selective use of radiation in intermediate risk rectal cancer. The co-primary endpoints of disease-free survival (DFS) and time to local recurrence (TLR) were assessed jointly by the sequential hypothesis testing procedure. DFS was tested for noninferiority first and then the superiority if the noninferiority was established. Then if the superiority test was inconclusive, the TLR was tested for noninferiority. This design can also incorporate different endpoints in phase II stage. The decision rules and study characteristics were optimized based on simulation studies.

Tolerance Limit For Longitudinal Data With Application In Establishing Withdrawal Period

✦Judy X. Li, Food and Drug Administration, 7500 Standish Place, MPN II, HFV-160, Rockville, MD 20855, *judy.li@fda.hhs.gov*; Oscar Chiesa, Food and Drug Administration; Min Zhu, SAS Institute Inc.

Key Words: tolerance limit, longitudinal data, withdrawal period

Withdrawal period is the interval between the time of the last administration of a compound and the time when the animal can be safely slaughtered for food. Tolerance limit is often calculated for establishing a withdrawal period. Tolerance limit within the context of longitudinal data is derived. The proposed tolerance intervals are validated with the simulation studies. We also illustrate the implementation with the data of Penicillin concentration in bovine kidney tissue.

Evaluation Of Different Numbers Of Dilution Serials Of Biological Samples Tested In Cell-Based Functional Assays

✦Liping Song, Merck & Co., Inc, 777 Sumneytown Pike, West Point, PA 19486 USA, *liping.song@merck.com*; Robert Capen, Merck & Co., Inc

Key Words: parallelism, relative potency, duplicate dilution serial

In cell-based functional assays, it is a common practice for the laboratory to test a biological sample along with a corresponding reference standard both in multiple dilution serials on 96-well micro-titer plates in at least three independent plates. The relative potency of the test sample is determined after confirming parallelism of corresponding dilution response curves from each plate, and the geometric mean relative potency from three independent plates is used to estimate the true relative potency of the sample. How many dilution serials of a sample and reference standard are sufficient on each plate: singlet, duplicate, triplicate, or even quadruplicate? Too few replicates might introduce an error in measurement, leading to an unreliable estimate. Too more replicates would be time consuming and waste resources. This presentation compares the determined relative potencies and associated variability with different numbers of dilution serials of a sample through bootstrap simulation. The results show that duplicate dilution serials of a sample are sufficient in most cases although multiple dilution serials provide some benefits during early development of the assays.

Quantification Of Impact Of Safety Monitoring On Type I Error And Power Of Efficacy Analysis In Phase Iii Group Sequential Clinical Trial

✦Yanqiu Weng, Division of Biostatistics and Epidemiology, Medical University of South Carolina, 135 Cannon St Suite 305R, Charleston, SC 29403, *ricerweng@gmail.com*; Wenle Zhao, Division of Biostatistics and Epidemiology, Medical University of South Carolina; Yuko Palesch, Division of Biostatistics and Epidemiology, Medical University of South Carolina

Key Words: phase III clinical trials, group sequential analysis, stopping guidelines, safety and efficacy, bivariate binary response, power

In large phase III clinical trials, group sequential analysis is a common statistical approach to monitor efficacy outcome during the trial. For some life-threatening diseases, safety is also monitored even more frequently than efficacy. However, interim analysis for overwhelming efficacy ignores multiple analyses of the safety outcome up to that point, and when safety and efficacy endpoints are highly correlated, how much does interim safety analyses cost on the type I and II error probabilities of the interim efficacy analyses? We use a multivariate normal approximation approach to estimate the probability of stopping for efficacy and/or safety for trials that concurrently monitor binary efficacy and safety endpoints. The estimation results are verified by simulation method. Our study suggests, type I error decreases as efficacy-safety correlation increases. Moreover, while power for efficacy is robust to the misspecification of the magnitude of safety-efficacy correlation, it is vulnerable to the variation of safety profile: power decreases significantly with between-group difference in safety probabilities increases, even if such difference is clinically trivial.

Reinforcement Learning Strategies For Clinical Trials In Non-Small Cell Lung Cancer

✦Yufan Zhao, Amgen Inc., One Amgen Drive, Thousand Oaks, CA 91320, *yufanz@amgen.com*

Key Words: reinforcement learning, Q-learning, support vector regression, clinical trials, personalized medicine, non-small cell lung cancer

We present a reinforcement learning design to discover optimal individualized treatment regimens for a non-small cell lung cancer trial. In addition to the complexity of the problem of selecting optimal compounds for first and second-line treatments based on prognostic factors, another primary scientific goal is to determine the optimal time to initiate second-line therapy, either immediately or delayed after induction therapy, yielding the longest overall survival time. Q-learning is utilized and approximating the Q-function with time-indexed parameters can be achieved by using support vector regressions. A simulation study shows that the procedure not only successfully identifies optimal strategies of two lines treatment from clinical data, but also reliably selects the best time to initial second-line therapy while taking into account heterogeneities of NSCLC across patients.

Evaluation Of Power And Type I Error On Different Statistical Methods Analyzing Hypoglycemia Data Using Bootstrap Simulation

✦Honghua Jiang, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana 46285 US, IN 46285, *jianghh@lilly.com*; William Huster, Eli Lilly and Company; Xiao Ni , Eli Lilly and Company

Key Words: hypoglycemia, bootstrap

Hypoglycemia has long been recognized as a major barrier to achieving normoglycemia with intensive diabetic therapy. It is one common safety concern for the diabetes patients. Therefore, the proper application of statistical methods for analyses of the hypoglycemia data is of importance. The Poisson model is commonly used to analyze count data, like the hypoglycemia event data. However, one characteristic of the Poisson distribution is that its mean and variance are identical. The negative binomial model is an alternative for over-dispersed count data in which the variance exceeds the mean. Non-parametric rank AN(C)OVA model is another way to analyze the hypoglycemia data. Sometimes, simple AN(C)OVA model is also used to analyze the hypoglyce-

mia data. Bootstrap simulation studies are conducted to evaluate the power and type I error of these four statistical methods based on the data from a diabetes clinical trial.

Estimating Controlled Direct Effects For Time-Varying Treatments Using Structural Nested Mean Models

✦Tomohiro Shinozaki, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0003 Japan, *shinozaki@epistat.m.u-tokyo.ac.jp*; Yutaka Matsuyama, The University of Tokyo; Yasuo Ohashi, The University of Tokyo

Key Words: direct effect, g-estimation, longitudinal data, structural nested mean model, time-dependent confounding

Estimating direct effects without bias requires that two assumptions hold, that is, the absence of unmeasured confounding for treatment and outcome, and the intermediate and outcome. Even if the above two assumptions hold, one cannot estimate direct effects via standard methods such as stratification or regression modeling if treatment affects confounding factors, namely, time-dependent confounders. Sequential g-estimation method for the structural nested mean models has been developed for estimating direct effects in the presence of time-dependent confounders. In this talk, we extend this method for data with time-varying treatments and repeated measured intermediates. Simulation studies showed that usual regression approaches were heavily biased in the presence of time-dependent confounders, but our sequential g-estimator remained unbiased. The proposed method was applied to data from a large primary prevention trial for coronary events in which plavastatin was used for lowering cholesterol. Our analyses showed that, for patients with moderate hypercholesterolemia, the benefit experienced by pravastatin could not be attributed to the effect of treatment on the cholesterol levels.

The Ratio Of Median Or Mean Change Scores For Treatment Comparisons

✦Joseph T Wang, Pfizer Inc., 235 East 42nd St., New York, NY 10017, *joseph.t.wang@pfizer.com*; Kelly H. Zou, Pfizer Inc; Martin Ove Carlsson, Pfizer Inc.

Key Words: Ratio statistic, Chang Score, Effect Size, Bootstrap, Normal-Ratio Distribution, Ratio T-Test

When comparing mean change scores of diary endpoints in clinical trials for treating subjects with overactive bladder (OAB) syndrome between two different groups, two-sample tests and the analysis of covariance with adjustments for baseline and additional covariates are frequently used to assess treatment effect. However, the reduction of diary endpoints in the OAB trials is often within a narrow range (e.g., -5 to 0), and a small mean change may not be a convincing treatment effect to physician/patient. When it is of interest to make a superiority claim of one treatment over another, the ratio of the mean or median change scores of two treatments is useful for the claim. The nonparametric bootstrap may be conducted. Alternatively, a normal-ratio distribution of the ratio statistic may be assumed and a parametric ratio t-test conducted. Finally, an extension to the ratio of two correlated random variables is briefly considered. The above methods for analyzing the ratio statistic are illustrated by pooling two identical randomized clinical trials in treating OAB. Monte-Carlo simulations are performed to compare the performances of these methods.

Statistical Analysis In Oc Pre-Treated Infertility Trials

✦Dixi Xue, Merck Inc. & Co., 1 E. Scott Ave, Rahway, NJ USA, *dixi.xue@spcorp.com*

Key Words: OC, GnRH agonist, recFSH, COS, meta-analysis

Oral contraceptives (OC) are quite frequently used in both GnRH agonist and GnRH antagonist regimens. Some centers use them and some don't. So, it is really hard to say which the current standard of care is. In the antagonist regimens the main reason for using them is to allow scheduling of the start of stimulation, the benefits of which are clearly shown in the article of Barmat et al (2005). One hypothesized benefit is that the OC pre-treatment would better synchronize the cohort of follicles possibly leading to a better stimulation. Our current study is to identify factor(s) capable of predicting ovarian response in women undergoing COS (Controlled Oocyte Stimulation) with a fixed daily dose of recFSH by multiple regression models and to find the OC/non-OC impact on ovarian response by meta-analysis.

Adaptive Clinical Trial Designs With Pre-Specified Rules For Modifying The Sample Size Based On The Interim Estimate Of Treatment Effect

✦Gregory Levin, University of Washington, , *glevin11@uw.edu*

Key Words: Adaptive designs, Clinical trials, Conditional power, Sample size modification, Group sequential tests, Sufficiency

Methods allowing unplanned interim adaptations to the sample size suffer losses in efficiency when compared to group sequential designs (Jennison and Turnbull, 2006, 2003; Tsiatis and Mehta, 2003). However, when adaptive sampling plans are completely pre-specified, inference can be based on the minimal sufficient statistic. In two general settings, we quantify the relative costs and benefits of a variety of fixed sample, group sequential, and pre-specified adaptive designs. We find symmetric pre-specified adaptive designs that are "optimal" in that they minimize the expected sample size at the design alternatives. Our results suggest that optimal pre-specified adaptive designs can lead to very small efficiency gains over optimal group sequential designs with the same number of analyses. We also describe optimal adaptation boundaries based on several different timings of the adaptation analysis and on a variety of different scales for the interim test statistic. These findings provide insight into what are good and bad choices of adaptive sampling plans and suggest that adaptive designs proposed in the literature are often based on inefficient rules for modifying the sample size.

Sample Size Calculations In Clinical Trials With Binary Data: Theoretical And Practical Issues

✦Arminda Siqueira, Universidade Federal de Minas Gerais (UFMG), International Brazil, *profa.arminda@yahoo.com.br*

Key Words: binary data, clinical trials, closed-form formula, sample size

Sample size, an important element in the planning stage of clinical trials, can be determined from closed-form formulas, in iterative equation solving procedures or via simulation. While closed-form formulas are practical and convenient, their accuracy varies, therefore simulated solutions are often preferred. For binary response trials, the sample size

calculation depends on many factors, such as type of study (superiority, non-inferiority and equivalence), design (parallel, crossover), measure (e.g. difference, odds ratio, relative risk), test (e.g. Wald, score and likelihood ratio), solution (exact, asymptotic) and statistical method (classical, Bayesian). This paper presents several comparative analyses of both analytical and simulation-based approaches, and discusses theoretical and practical issues regarding the appropriate methods of calculating sample size for clinical trials with binary response.

424 Contributed Oral Poster Presentations: Section on Bayesian Statistical Science

Section on Bayesian Statistical Science

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Sample Size Determination for Drug Combination in Fixed Dose Trials

✦Jie Wang, Quintiles. Inc, 6700 W 115th Street, Overland Park, KS 66211, *jie.wang@quintiles.com*

Key Words: sample size estimation, Bayesian, drug combination

We have considered the problem of drug combination in fixed dose trials to test whether a drug mixture, which may combine two or more agents, is more 'effective' than each of its components. Informative priors are derived for component drugs and a non-informative prior is assumed for the drug mixture. Sample sizes are evaluated by posterior standard errors, average probability of more effectiveness and Bayesian power.

Coherence In Hypothesis Testing

✦Rafael Izbicki, Carnegie Mellon University, 3401 Forbes Avenue Apt 104, Pittsburgh, PA (PIT), PA 15213 US, *rafaelizbicki@gmail.com*; Luìs Gustavo Esteves, Universidade de Sao Paulo

Key Words: Simultaneous tests procedures, Decision Theory, Decision rules, Monotonicity

In Statistical Inference, it is usual to simultaneously test a set of hypotheses concerning a parameter. Our objective is to evaluate the (lack of) logical coherence among conclusions obtained from tests conducted regarding the same data. In this study, we present a definition of class of hypotheses tests, a function that associates a test function to each hypothesis of interest. Some properties that reflect what one could expect (in terms of logical coherence) from tests for different hypotheses are then evaluated. These properties are exemplified by classes of hypotheses tests that respect them and, whenever possible, justified on a decision-theoretic framework. Then, based on the properties studied, we propose sets of axioms regarding classes of hypotheses tests. Usual classes of hypotheses tests (e.g. Bayesian tests based on posterior probabilities, likelihood ratio tests) are inve

Bayesian Estimation Of Logistic Regression With Misclassified Covariates And Response For Educational Psychology Data

✦Brandi Falley, Baylor University, , *brandi_falley@baylor.edu*; James Stamey, Baylor University; Alex Beaujean, Baylor University

Key Words: Bayesian Estimation, Logistic Regression, Misclassified Covariates, Educational Psychology

Measurement error problems in binary regression are of considerable interest among researchers, especially in epidemiological studies. Misclassification can be considered a special case of measurement error specifically for the situation when measurement is the categorical classification of items. Bayesian methods offer practical advantages for the analysis of epidemiological data including the possibility of incorporating relevant prior scientific information and the ability to make inferences that do not rely on large sample assumptions. In this presentation we consider a logistic regression model where both the response and a binary covariate are subject to misclassification. We assume both a continuous measure and a binary diganostic test are available for the response variable but no gold standard test is assumed available. We consider a fully Bayesian analysis that affords such adjustments, accounting for the sources of error and correcting estimates of the regression parameters.

Modeling Response Errors In Repeated Self-Report Surveys: An Example Of Multiple Editing For Categorical Data

✦Robin Jeffries, University of California, Los Angeles, Los Angeles, CA 90034, *rjeffries@ucla.edu*; Robert E. Weiss, University of California, Los Angeles

Key Words: multiple editing, survey, multiple imputation, bayesian, adolescents, self-report

Conflicting answers to survey questions, and inconsistent responses to identical repeated questions such as gender, exist in longitudinal self-report surveys. Deterministic data editing techniques correct these errors but subsequent analysis assumes the edit is correct and does not allow for edit error. We propose models to perform multiple edit in direct analogy with multiple imputation. We multiply edit erroneous data under a model and combine the multiply edited data sets using Rubin's rules for combining multiply imputed data sets. This requires a model for the missing correct data given the clearly incorrect responses. We illustrate this process by considering a Bayesian latent variable model for student reports of being born in the US and how that varies as a function of age and ethnicity. We illustrate a conditional probit model of cell probabilities for contingency tables with what should be a structural zero and apply it to model the probability the student knows about and uses a condom distribution program on campus. The motivating data set consists of a four year sample from a longitudinal intervention study on Los Angeles middle- and high-school students.

Bayesian Model Checking In Hierarchical Spatial Models For Count Data

✦Liang Jing, University of Texas at San Antonio, One UTSA Circle, Department of MSS, San Antonio, TX 78249-0631, *liang.jing@utsa.edu*; Victor De Oliveira , University of Texas at San Antonio

Key Words: model checking, latent process, incompatibility, transformed residual

Hierarchical spatial models for count data are increasingly used for data analysis in many earth sciences, but model checking and model selection in this class of models remain difficult tasks due to the presence of an unobservable latent process. For this class of models, we investigate the application of model checking methods based on measures of relative predictive surprise, as those described in Bayarri and Castellanos (2007). We also propose an alternative method to diagnose incompatibility between model and data based on a kind of transformed residuals. The usefulness of the proposed model checking methods is explored using both simulated and real spatial count data.

Modeling Temporal Dynamics Of Vaginal Bacterial Communities In Healthy Woman

✦Zaid Abdo, University of Idaho, 83844 USA, *zabdo@uidaho.edu*; Ursel Schutte, Indiana University

Key Words: Hierarchical models, Bayesian Statistics, Mixture models

The beneficial effects of the endogenous microbiota on women's health are numerous and poorly characterized. Significant alterations or disruptions of the vaginal microbiome, such as in bacterial vaginosis, may increase women's risk including preterm labor and acquisition of infections. We developed a hierarchical Bayesian framework in an effort to understand factors associated with maintenance and fluctuation of the vaginal microbiota over time. Our framework includes modeling changes in the microbial community structure using a Bernoulli distribution to describe presence and absence of each phylotype and a Poisson distribution to describe their abundances. Metadata--environmental, behavioral, and life history data-were linked using a generalized linear model. This approach allowed for detecting dependencies between bacterial community structures and observed metadata. Preliminary results suggest that the temporal dynamics of bacterial communities are personalized, that is there are no obvious patterns in changes in vaginal bacterial communities over time among woman but rather the dynamics are individual to each woman.

The Valence Bond - A Bayesian Approach To Patient Linking

Brandon Barber, Valence Health; ✦Bart Phillips, Valence Health, 600 W Jackson Blvd, Ste. #800, Chicago, IL 60661, *bphillips@valencehealth.com*

Key Words: Linking Algorithm, Bayes, Member Identification, Clinical Integration

As US healthcare increasingly stresses data sharing and interoperability, the need to accurately and efficiently link patients from disparate data sources will intensify. Use of social security numbers as means of medical identification, however, is a dwindling trend and without an unambiguous patient identifier, challenges will abound regarding patient linkage. We propose that a patient identification system reliant on common claim-driven information and independent of SSN will meet our needs to link records over sources such as laboratory, hospital, and practice management systems. Our Bayesian approach to probabilistic linking matches claim lines based on a user-defined degree of certainty and a minimum amount of blocking for computation practicability. We support our method through a series of posterior distribution analyses based on potential user-defined degrees of certainty.

Interval-Censored Negative Binomial Regression

✦Stephanie Simon, Baylor University, 66 Daughtrey Ave, Apt 215, Waco, TX 76706, *stephanie_simon@baylor.edu*

Key Words: Interval-censored, negative binomial regression, survey, Bayesian

Interval-censored counts arise in various applications, including survey data in which the respondent is questioned about a sensitive subject, such as number of sexual partners, or asked to recall the number of occurrences of some event, such as headaches in the last week. Suppose the true count falls in the interval [j,k] with probability 1 but that we only observe the lower and upper bounds of that interval, j and k. We consider a negative binomial regression on such interval-censored counts. We take a Bayesian approach in our analysis, and in a simulation study we compare our results with those obtained from similar analysis when the true counts are known.

Analysis Of Triangulation Methods In Bayesian Networks

✦Chansoo Kim, Kongju National University, Dep of Applied Mathematic, Kongju National Univ, 182 SHINKWAN-Dong, Kong-ju, 314-701 Republic of Korea, *chanskim@kongju.ac.kr*; Dong Hoon Lim, Gyeongsang National University

Key Words: Bayesian Networks, Genetic algorithm, Optimal decomposition, Triangulation

A Bayesian network is a graphical model for probabilistic relationships among a set of variables. The search for an optimal node elimination sequence for the triangulation of Bayesian networks is an NP-hard problem. In this paper, we examine the most frequently used optimality criteria and consider the optimal decomposition of Bayesian networks using genetic algorithms with various genetic operators.

Incomplete Data As A Predictor In Multiple Imputation Models

✦Tracy Pondo, CDC, MS C25, 1600 Clifton Rd NE, Atlanta, GA 30329, *TPondo@cdc.gov*; Elizabeth Zell, CDC; Melissa M Lewis, CDC

Key Words: Multiple Imputation, Surveillance, Disease, Race

Multiple imputation of all missing variables in a data set can be performed by sequential regression multivariate imputation to create a complete data set for statistical analysis. Imputations are performed on all variables in the imputation model that have missing values. The imputed values of one variable are used as predictors to impute the values of other variables. An alternative to imputation of all categorical variables in the imputation model is the creation of new categories to represent missing data. To demonstrate the advantages and drawbacks of using incomplete data as a predictor in the imputation model we will compare imputed data sets with and without missing data categories in the imputation model.

Latent Class Analysis Of Multivariate Longitudinal Data: Uncovering Response Patterns In A Stem Cell Study

✦Qianqiu Li, Johnson & Johnson, PA 19087, *qianqiu.li@yahoo.com*; Xiaozhen Wang, Johnson & Johnson; Ian Harris, Johnson & Johnson; Bill Pikounis, Johnson & Johnson

Key Words: Latent class, Mixture model, Longitudinal data, Small sample, Response Pattern, Stem Cell

By applying latent class analysis to longitudinal data from a stem cell hCTC study with 3 time points and 5 treatment groups, we distinguished 3- and 4- subgroup mean response profiles and revealed their association with treatment and time. These results indicate efficacy of two dosed groups over 12 weeks, compared with only short-term (or 4-week) improvement for a comparator treatment group, and no improvement in the Vehicle group. In comparison with ANOVA and mixed model analyses, latent class analysis can be a powerful tool for inference on the mean response profile as well as posterior probabilities for membership of response subgroups. A maximum likelihood approach via the EM algorithm is used. The results demonstrate that the latent model is particularly effective in analysis of small sample data from animal models in presence of nonnegligible inter-individual variability.

In Defense Of Randomization: A Subjectivist Bayesian Approach

Fernando Vieira Bonassi, Department of Statistical Science - Duke Univeristy; ✦Raphael Nishimura, Survey Methodology Program, University of Michigan, 426 Thompson Street, Room 4050, Ann Arbor, MI 48104, *raphaeln@umich.edu*; Rafael Bassi Stern, Department of Statistics, Carnegie Mellon University

Key Words: Bayesian Statistics, Decision Theory, Game Theory, Intersubjectivism, Randomization

In research situations usually approached by Decision Theory, it is only considered one researcher who collects a sample and makes a decision based on it. It can be shown that randomization of the sample does not improve the utility of the obtained results. Nevertheless, we present situations in which this approach is not satisfactory. First, we present a case in which randomization can be an important tool in order to achieve agreement between people with different opinions. Next, we present another situation in which there are two agents: the researcher - a person who collects the sample; and the decision-maker - a person who makes decisions based on the sample collected. We show that problems emerge when the decision-maker allows the researcher to arbitrarily choose a sample. We also show that the decision-maker maximizes his expected utility requiring that the sample is collected randomly.

Bayesian Surface Smoothing Under Anisotropy For Count Data

✦Subhashish Chakravarty, University of Southern California, 10604 Wilkins Avenue, #302, Los Angeles, CA 90024, *subhashc@marshall.usc.edu*

Key Words: Bayesian, smoothing, anisotropy, count data

We propose a non-parametric approach to Bayesian surface smoothing for count data in the presence of geometric anisotropy. We use eigenfunctions generated by thin-plate splines as the basis functions of the smooth surface. Using eigenfunctions does away with having to place knots arbitrarily and the non-parametric approach provides for modeling flexibility. The smoothing parameter, the anisotropy matrix, and other parameters are simultaneously updated by a Reversible Jump Markov Chain Monte Carlo (RJMCMC) sampler. Model selection is done concurrently with the parameter updates. Since the posterior distribution of the coefficients of the basis functions for any given model is available in closed form, we are able to simplify the sampling algorithm in the model selection step. We find higher values of the smoothness parameter correspond to more number of basis functions being selected. A Bayesian approach also allows us to include the results obtained from previous analysis of the same data, if any, as prior information. It also allows us to evaluate point-wise estimates of variability of the fitted surface.

Bayesian Modelling Of Compositional Time Series Data

✦Gabriela Czanner, Department of Statistics, University of Warwick, Coventry, International CV4&AL United Kingdom, *Gabriela.Czanner@warwick.ac.uk*; Jim Q. Smith, Department of Statistics; John Fenlon, Department of Statistics

Key Words: Dynamic Linear Models, Bayesian Time Series, Aitchison geometry

We consider Bayesian Time Series models (Dynamic Linear Models, Harrison and West, 1997) for analysis of time series of compositional data. Traditionally, the key to analysing compositional data is to look at the relative magnitudes and variations of components, rather than their absolute values. This is done by using log-ratios of components of compositions. Consequently the whole D-dimensional supply space is projected onto a D-1 dimensional space (Aitchison, 1986). Here, we discuss an alternative approach where data are modelled in their original scale while imposing the restrictions thus leading to a constrained linear dynamic model. Then, as a novel approach, we use the constrained linear dynamic model to impose the probability distribution on the D-1 dimensional space of compositions. The primary use of these methods is in supply-chain management and demand forecasting. We illustrate the methods in daily gas supply data which exhibit the structural changes, an increasing uncertainty, the constraint that total supply must equal demand; and the physical constraints of the gas transmission system.

A Bayesian Decision-Based Model For Aggregating Expert'S Information

✦MarÌa Jes´s Rufo, University of Extremadura Escuela PolitÈcnica, Avda de la Universidad s/n, C·ceres, 10071 Spain, *mrufo@unex.es*; Jacinto MartÌn, University of Extremadura Facultad de Ciencias; Carlos Javier PÈrez, University of Extremadura. Facultad de Veterinaria

Key Words: Bayesian decision model, Kullback-Leibler divergence, Opinion pooling

This work provides a decision-based approach to assess the weights in a logarithmic pooling of prior distributions. Each expert provides prior information over the quantity of interest as a proper prior distribution.

Then, the decision maker combines them through a logarithmic pooling. Next, the weights have to be assessed to obtain the full aggregated prior distribution. In order to do it, the problem is formulated as a decision one. Therefore, given the decision space and the states of nature, an appropriated loss function based on Kullback-Leibler divergence is defined. Two situations are considered depending on whether the decision maker assumes prior ignorance about the quantity of interest or not. These situations are distinguished through the choice of suitable prior distributions over the state of nature. Several methods are considered in order to obtain this prior distribution. Hence, the optimal weights are those that minimize the expected loss. Finally, the results obtained under the two considered frames are compared.

Seqbayes: An Adaptive Bayesian Framework For Calling Genotypes From Next-Generation Sequence Data

✦Daniel D Kinnamon, University of Miami Miller School of Medicine, Department of Human Genetics, P.O. Box 019132 (M-860), Miami, FL 33101, *dkinnamon@med.miami.edu*; Eric H Powell, University of Miami Miller School of Medicine; Michael A Schmidt, University of Miami Miller School of Medicine; Eden R Martin, University of Miami Miller School of Medicine

Key Words: genotype calling, next-generation sequencing, Markov chain Monte Carlo

Calling individual genotypes from next-generation sequence data requires estimating their posterior probabilities using a joint model for the observed nucleotide read data and latent genotypes. Existing approaches estimate posterior probabilities by substituting either fixed parameter values based on prior knowledge or MLEs obtained using the current sample of independent individuals into this model. While substituting MLEs yields lower average genotype-call error rates, some small samples with sparse information may have higher genotype-call error rates due to non-identifiability. We propose a Bayesian genotype-calling approach that solves this problem by assigning priors to all parameters in the model. These priors can improve identifiability by combining existing knowledge with current sample information in an adaptive manner. We also propose an MCMC algorithm for estimating marginal posterior genotype probabilities under our Bayesian model. In simulations, we show that our MCMC algorithm has favorable empirical properties. We also demonstrate that our approach yields lower average and worst-case genotype-call error rates than using MLEs in small samples with sparse information.

Bayesian Inference For Mean Residual Life Functions In Survival Analysis

✦Valerie Poynor, University of California, Santa Cruz, , *vpoynor@soe.ucsc.edu*

Key Words: bayesian nonparametrics, survival analysis, Dirichlet Process, mean residual life, right censoring

In survival analysis interest lies in modeling data that describe the time to a particular event. Informative functions, namely the hazard function and mean residual life function (MRL) , can be obtained from the model's distribution function. The MRL function provides the expected remaining life given survival up to a particular time. This function is of interest in reliability, medical, and actuarial fields. The MRL function not only has a simple and practical interpretation, it characterizes the distribution through the Inversion Formula. Thus the MRL function can be used in fitting a model to the data. We review the key properties of the MRL function and investigate its form for some common distributions. We also study Bayesian nonparametric inference for MRL functions obtained from a flexible mixture model for the corresponding survival distribution. In particular, we develop Markov Chain Monte Carlo posterior simulation methods to fit a nonparametric lognormal Dirichlet process mixture model to two experimental groups. We perform a model comparison with a parametric exponentiated Weibull model. Finally, we fit a nonparametric mixture model to a right censored dataset.

Bayesian Elastic Net For Multi-Class Classification And Survival Analysis

✦Lingling Zheng, Duke University, durham, NC , *lz35@duke.edu*

Key Words: Variational Bayesian, Bayesian Elastic Net, Survival Analysis, Multinomial Probit Regression, Variable Selection, Gibbs Sampling

Bayesian Elastic Net is an advanced technique for addressing the problem of grouped variable selection and sparseness.I am interested in adopting this strategy for clinical application. In this paper, I develop a multinomial probit regression model of Bayesian Elastic Net for classification problem, i.e. identifying classifiers to infer a set of important and highly correlated predictors, e.g. genes or peptides. Additionally, in order to study association between survival and gene expression signature, I present censored exponential regression model of Bayesian Elastic Net. Furthermore, missing data imputation is also considered for both cases. Inference of these approaches is conducted through both Gibbs sampling and variational Bayesian (VB) approximation. The two models are validated by first performing simulation on toy datasets; then I obtain biochemical measurements data from sepsis patients to classify and predict their disease status. Finally, I assess Bayesian survival model on lung cancer patients' microarray data and failure time. The methods show that Gibbs sampling has better accuracy in classification, while VB tends to achieve better sparseness and efficiency.

Bayesian Model Selection Using Non-Local Priors

✦Lu Wang, University of California, Los Angeles, , *lu.wang@ucla.edu*

Key Words: Bayesian model selection, non-local priors, Gibbs sampling

Bayesian model selection method is applied for smoothing splines through knot selection. We compared different priors on knot selection problem and found corresponding shrinkage penalization for each prior. We also applied non-local priors on model selection for generalized linear models. Furthermore, computational issues for posterior sampling was addressed.

A Graph Theoretic Analysis Of Brain Connectivity In Schizophrenia Patients Using A Bayesian Hierarchical Model

✦Lijun Zhang, Emory University, Dept. of Biostatistics and Bioinformatics, 1518 Clifton Rd, NE, Atlanta, GA 30322, *l.zhang@emory.edu*; DuBois Bowman, Emory University

Key Words: Functional connectivity, graph theoretic, Bayesian hierarchical model

Functional connectivity in the brain reflects associations in neural activity between distinct brain regions and has been actively studied in healthy subjects as well as in patients with psychiatric disorders. Recently, graph theoretic metrics have been applied to characterize functional networks, addressing properties such as centrality, clustering, efficiency, and small world. We propose a Bayesian framework for obtaining group-level estimates to build graph theoretic properties of task-related functional networks. Moreover, our approach enables inferences about group comparisons in functional connectivity properties. We apply our methods to an fMRI study to evaluate differences in working-memory related functional connectivity between schizophrenia patients and healthy controls.

425 Contributed Oral Poster Presentations: Section on Risk Analysis

Section on Risk Analysis

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Concentration-Response Estimation For High -Throughput Screening Data

✦Rhyne Woodrow Setzer, National Center for Computational Toxicology, ORD, US EPA, Research Triangle Park, NC , *setzer.woodrow@epa.gov*

Key Words: High-throughput, ToxCastT, concentration-response, P-splines, micro-titer, Bayesian

Toxicologists and risk assessors are exploring the use in environmental risk assessment of data from high-throughput screens in which multiple (hundreds or thousands) chemicals are tested in 96-well or larger micro-titer plates for activity against multiple (dozens to hundreds) assays in concentration-response designs. Assays are conducted using laboratory robots, and chemicals and concentrations are applied in systematic patterns on plates. The use of automation can lead to both systematic plate effects and variability best modeled as autocorrelated errors. This presentation discusses evaluation of systematic plate effects and estimation of concentration-response curves in high-throughput data taking into account both systematic and auto correlated errors, illustrated using assays from the ToxCastT project of the USEPA. P-splines estimated using Bayesian methods are used to adjust out plate effects and estimate concentration-response curves, and to estimate concentrations associated with specified changes from background for each chemicalXassay combination. Uncertainties are characterized with posterior intervals. This abstract does not necessarily reflect U.S. EPA policy.

426 Contributed Oral Poster Presentations: Section on Statistics in Defense and National Security

Section on Statistics in Defense and National Security

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Anomaly Detection And Data Fusion Techniques Applied To Radiation Portal Monitor Data

✦Dov Cohen, Sandia National Laboratories, 7011 East Ave, Livermore, CA 94551 USA, *idcohen@sandia.gov*; Isaac Shokair, Sandia National Laboratories

Key Words: Anomaly Detection, Data Fusion, Radiological Data, Gamma-ray, Neutron

Radiation portal monitors (RPMs) are used to screen vehicles and containers for the presence of radiological threats. To maximize sensitivity to weak signals and minimize deployment costs, RPMs use poly-vinyl toluene (PVT) detectors which have large geometric collection efficiencies but poor energy resolution. Since naturally occurring radioactive materials (NORM) are ubiquitous in commerce, the effectiveness of RPM detectors depends on their ability to differentiate between benign and threat sources. Because encounters with threat sources are rare events, statistical anomaly detection algorithms may improve the power of classifiers that distinguish benign and threat sources. Such classifiers compare single measurements with historical measurements of benign sources. The challenge is to determine objective ways of detecting anomalies in radiation data. Exploratory data analysis techniques and unsupervised machine learning are used to model the multimodal structure of the benign datasets. In this paper we will describe anomaly detection using statistical distance measures, and the use of data-fusion techniques to combine signals from disparate detectors (i.e. gamma-ray and neutron).

427 Contributed Oral Poster Presentations: Section on Survey Research Methods

Section on Survey Research Methods

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

The Impact of Visual Design in Survey Cover Letters on Response and Web Take-Up Rates

✦William Mockovak, Bureau of Labor Statistics, 2 Massachusetts Ave, N.E., Suite 1950/OSMR, Washington, DC 20212, *Mockovak_W@BLS.gov*

Key Words: Cover letter, Web take-up rate, Information mapping, Visual design

Cover letters for mailed survey forms can differ in a variety of ways. Previous research suggests that visual design can impact response, and that the effects might even be negative. Therefore, the purpose of this study was to continue this line of research and investigate the impact of a unique visual design for a cover letter, while holding the content constant. Whereas the possible types of visual design changes are numerous, this study looked at the impact of only one type of approach (Information Mapping) that has been shown to improve the usability of written materials in other forms of communication. A sample of 1,000 addresses was randomly assigned to either an experimental or control condition (500 in each), and the response rate was analyzed after one mailing attempt. Results showed no statistically significant differences between groups in response rates (overall response was 27.6

percent). A secondary objective was to determine how many respondents would opt to use a simple Web-reporting option when one was offered in the cover letter. Only 2.5 percent chose the Web option, with ten times as many choosing to respond using the mail, and no differences between groups.

Comparison of Variance Estimates in a National Health Survey

✦Karen E Davis, Agency for Healthcare Research and Quality, 540 Gaither Road, Room 5355, Rockville, MD 20850, *karen.davis@ahrq.gov*; Van Parsons, National Center for Health Statistics

Key Words: Sample survey, Variance estimation

The National Health Interview Survey (NHIS) is one of the major data collection programs of the National Center for Health Statistics (NCHS). This survey has a complex design that covers an approximately 10-year sample design period, and was redesigned most recently with the 2006 NHIS. The actual design features multiple stages of sampling and weighting adjustments. Simplified and user-friendly procedures have been developed for both in-house and public-use design-based analyses for use with linearization-based methods. In particular, for NHIS public-use data, a standardized design, modified to prevent identification of sampled geographical areas and simplified to consist of two sampled clusters per stratum, has been provided. For simplified procedures to be acceptable to the NHIS-user community, estimates of standard errors produced from simplified structures should be close to those produced by using detailed NHIS design and weighting information, e.g., using Yates-Grundy-Sen forms for two-stage variances and linearization of the final weights. In this study the standard error estimates produced from standardized designs are compared to those produced using detailed methods

Nonresponse Bias In The Survey Of Youth Perception Of Science And Technology In Bogota

✦Edgar Mauricio Bueno Castellanos, OCyT, Carrera 15 No 37 - 59, Bogota, International 111311 Colombia, *embuenoc@gmail.com*

Key Words: sampling design, nonresponse bias, calibration

The Colombian Observatory of Science and Technology -OCyT- developed, in 2009, a survey about the perception of Science and Technology in students of the last two years of high school in Bogot-, Colombia. The survey sampling design was stratified according to the nature of school. Two main sources of Nonresponse were detected. The first one, as a consequence of the difference in the response probability according to the nature of school: the survey was implemented in 15 out of the 16 official schools included in the original sample, while only 13 out of 31 private schools allowed to collect information. The second source corresponds to students who did not assist during the days when survey was applied. Estimates, initially, were obtained modifying the original sample sizes by those observed. Subsequently, it was decided to obtain new estimates taking into account the nonresponse effect; to achieve this goal, the values corresponding to item nonresponse were imputed using the methodology of the nearest neighbor and the calibration method was used for unit nonresponse. The results obtained for both cases don't show visible differences, especially when estimating a ratio.

Combined Methods For Imputing School Variables In Principal Data Files

✦yan wang, American Institutes for Research, 1990 K street, NW. suite 500, washington, DC 20006, *ywang@air.org*; Matthew Doyle, American Institutes for Research

Key Words: School and Staffing Survey, cold-deck imputation, regression impuation, school nonresponse, data utility, principal data files

Due to school nonresponse, three school variables-school level, school enrollment, and urbanicity-have about 10 percent missing values in the 1987-88, 1990-91, and 1993-94 School and Staffing Survey principal data files. To create fully imputed files to meeting reporting standards on key variables, a combination of imputation procedures were used for these variables based on the availability of auxiliary information. The first imputation method was to use existing sampling variables - the school stratum codes for public and private schools to extract school level values. The second method was a cold-deck imputation using universe data (Common Core Data school files and Private School Universe Survey data) to fill in school enrollment values when possible. Thirdly, as a last resort, a set of survey variables was selected to predict imputed variables and regression imputation was used to fill the rest of the missing values on school enrollment and urbanicity. The utility of the imputed data was evaluated by comparing the original data files with the imputed files on the weighted cell counts, standard errors, and pairwise correlations and multivariate associations.

Estimating The Bias Resulting From The Exclusion Of Cell Phone Only Respondents

Burton Levine, RTI International; ✦Christine Davies, RTI International, 3040 E. Cornwallis Rd, RTP, NC 27709, *cdavies@rti.org*; Bonnie Shook-Sa, RTI International

Key Words: RDD, survey, cell phones, dual-frame

More than 30% of the population has zero probability of selection in most landline RDD surveys due to the exclusion of cell phone only households, zero banks, and households with no phone service. This coverage error is problematic since methodological research indicates that cell phone only respondents are very different than landline respondents; thereby, converting coverage error into coverage bias. To mitigate the coverage error many telephone surveys have moved to a dual-frame design consisting of both cell phone and landline numbers. For surveys that implement a repeated cross-sectional design, comparing estimates of population parameters before and after the change from a landline RDD to a dual-frame RDD can be problematic because differences in estimates might be a consequence of the different sampling frames. This paper presents methodology used to adjust for bias caused by the exclusion of cell phone only respondents; thereby, enabling the comparison of estimates from repeated cross-sectional studies that change methodology from exclusively sampling from landlines to sampling from both landlines and cell phones.

Using Isotonic Regression To Estimate Order-Restricted Health Indicators For The 1997-2006 National Health Interview Survey

✦Van Parsons, National Center for Health Statistics, Metro IV Building Rm 3279, 3311 Toledo Road, Hyattsville, MD 20782, *vlp1@cdc.gov*

Key Words: monotonic estimation, complex survey, replication

Estimates of health measures within a population can often be modeled by order-restricted population parameters. Population subdomains based on gender, race or a time period often define order relations for selected disease prevalence among the subdomains, for example, interventions have decreased smoking prevalence over time. With complex survey data such estimation and inferences are typically made using design-based (general) linear models or direct-estimate cell-mean models. These models are easy to implement, but have possible drawbacks including linear model distortion of the true monotonic nature or violation of hypothesized orderings by the direct estimate orderings when using a "small sample" cell-means model. As an alternative to the standard methods, isotonic regression techniques can be used as a basis for order-preserving estimation and inference. These techniques avoid the drawbacks of the standard methods mentioned. Data from the 1997-2006 National Health Interview Survey will be used to estimate parameters for hypothesized total and partial orderings for select health statistics. Comparisons of the standard and isotonic methods will be presented and discussed.

On Estmatinig Two Sensitive Characteristics

✦Cheon-Sig Lee, Texas A&M University-Kingsville, Department of Mathematics, 700 University Blvd., Kingsville, TX 78363, *choensig@gmail.com*; Stephen A. Sedory, Texas A&M university-Kingsville; Sarjinder Singh, Texas A & M University - Kingsville

Key Words: Sensitive characteristics, Privacy guaranteed, Efficient models, Respondent's coopeartions, Reducing false responses, bias

The randomized response technique (RRT) is useful for reducing response error problems when potentially "Sensitive Questions" are present in surveys of human populations. Direct Questioning often results in either refusal or falsification of respondents' responses. Warner (1965) was the first to introduce the use of a randomization device to resolve such an issue in human populations by considering the problem of estimating the proportion of a single sensitive characteristic. In this paper, we consider the problem of the estimation of the proportions of two sensitive characteristics, say A, B and their intersection by making use of a new randomization device. Estimators based on two different models, simple and crossed, are developed and investigated from the efficiency point of views.

Overcoming Challenges To Sample Design In Iraq

✦David Peng, D3 Systems, Inc., 1209 North Taft Street, Unit K, Arlington, VA 22201 USA, *david.j.peng@d3systems.com*

Key Words: Sampling, Iraq, Post-conflict, Satellite, International

Iraq is an extremely difficult environment for conducting survey research. Security concerns coupled with the lack of recent census population parameters present difficult challenges to sample design. Obviously, there is a need for accurate sampling in Iraq as policy-makers and analysts attempt to shape Iraq's future. D3 Systems is currently exploring innovative approaches in sample design and implementation to overcome these challenges in Iraq. Currently, D3 uses the statistical data from a 2005 study conducted by the Iraqi Central Statistical Office (CSO) of the Iraqi Ministry of Planning as the base population framework for its national surveys of Iraq. In the absence of a new census, these estimates from 2005 are widely accepted as the best source of data for the distribution of survey respondents. However, the lack of up-to-date, consistent, verifiable sample frame raises issues about the representativeness of any sample. D3 proposes a new sample design that incorporates the use of satellite imagery technology. This technology, combined with ground-level knowledge, is used to map Iraq beyond the country's 18 provinces, 102 districts, and 232 sub-districts.

Effect Of Population Skewness On Variance Estimation For Pps Sampling

✦Chin-Fang Weng, University of Maryland, Statistics Program, Mathematics Department, College Park, MD 20742, *cfweng@umd.edu*

Key Words: variance estimation, PPS sampling, Hanson-Hurwitz, effect of skewness, Hajek estimator, Sen-Yates-Grundy

A finite population of size N is sampled without replacement with inclusion probabilities proportional to a size variable x, which is correlated with the variable of interest y. The population total of y is estimated by T, the Horvitz-Thompson estimator. The Sen-Yates-Grundy estimator of Var[T] is unbiased but may be negative. Also the Sen-Yates-Grundy estimator involves the two-unit inclusion probabilities, which are often unknown and which make evaluating the estimator cumbersome in large samples. Alternative variance estimators were proposed by Hanson and Hurwitz (1943), Hajek (1964), and Brewer and Donadio (2003). We compare the performance of these estimators when the distribution of the size variable x is highly skewed using simulation methods. The results suggest that high skewness is associated with poor reliability of these estimators.

Analyzing High-Dimensional Longitudinal Incomplete Data With Both Continuous And Ordinal Variables

✦Xiang Lu, UCLA Biostatistics, Room 51-267 CHS, Department of Biostatistics, Los Angeles, CA 90095, *xianglv77@gmail.com*

Key Words: multiple imputation, longitudinal, high-dimension, bayesian

This report outlines a Bayesian method using factor analysis for performing multiple imputation of missing values in multivariate longitudinal data with continuous and ordinal variables.The model accommodates the longitudinal structure of the outcome high-dimensional longitudinal variables by assuming they are explained by a reasonably small number of factors as well as by incorporating a linear model with random-effect structure so that estimates are more stable than would be produced by estimating every component of the covariance matrix. A Markov-chain Monte Carlo (MCMC) procedure is used to draw samples from posterior predictive distribution of the missing values. Comparison between the newly developed method and some existing multiple imputation methods are planned based on simulated data sets, and data from a study comparing two oral surgery procedures is also used to illustrate the properties of the proposed method.

An Overview Over Following Rules In Household Panels And Their Effect On Sample Size

✦Matthias Schonlau, RAND Corporation, Pittsburgh, PA , *matt@rand.org*; Nicole Watson, University of Melbourne; Martin Kroh, German Institute for Economic Research (DIW)

Key Words: survey, household panel, following rule, sample size

In household panels, typically all household members are surveyed. Because household composition changes over time, so-called following rules are implemented to decide whether to continue surveying household members who leave the household (e.g. former spouses/partners, grown children) in subsequent waves. Following rules have been largely ignored in the literature leaving panel designers unaware of the breadth of their options and forcing them to make ad hoc decisions. In particular, to what extent various following rules affect sample size over time is unknown. Such knowledge is important because sample size greatly affects costs. We find that household survey panels implement a wide variety of following rules but their effect on sample size is relatively limited.

Day-Of-Week, Time-Of-Year, And Meteorological Effects On Items On A Local Health Survey

✦Kevin J Konty, New York City Department of Health and Mental Hygiene, 125 worth st, rm 315 cn6, new york, NY 11215 usa, *konty@yahoo.com*

Key Words: survey quality, validity, health surveys, mental health, physical activity and nutrition, public use microdata

The New York City (NYC) Community Health Survey (CHS) is a telephone survey of adults conducted annually since 2002 by NYC's Department of Health and Mental Hygiene. The CHS collects approximately 9,500 interviews a year describing a wide variety of health outcomes and behaviors and is extensively used in public health planning and evaluation. A recent study of the items of the Kessler-6 non-specific psychological distress scale observed a distinct day-of-week pattern. Because the CHS is collected fairly evenly across days there is minimal impact on estimated trends. However, both variance calculations and measures of association with other variables may be affected if they also exhibit day-of-week patterns. To investigate this, we quantified day-of-week, time-of-year, and meteorological effects on various survey items including health status, physical activity, fruit and vegetable consumption, and drinking and sex behaviors. Though effects tend to be mild, they may alter estimates especially for smaller subpopulations suggesting a need to explicitly address the issue and this may have important ramifications for public-use microdata releases that do not include such information.

Using Discriminant Analysis To Select Model: Application In Production Control

Maria Emilia Camargo, Universidade de Santa Cruz do Sul; ✦Suzana Leit„o Russo, Federal University of Sergipe, Aracaju - SE, 49035490 Brazil, *suzana.ufs@hotmail.com*; Angela Isabel dos Santos Dullius, Universidade Federal de Santa Maria; Eric Dorion, Universidade de Caxias do Sul

Key Words: Forecast Model, Discriminant Analysis, Production Control

When a large number of time series are to be forecast on a regular basis, as in large scale inventory management or production control, the appropriate choice of a forecast model is important as it has the potential for large cost savings through improved accuracy. A possible solution to this problem is to select one best forecast model for all the series in the dataset. In this paper we using discriminant analysis to select one model representing for the daily production process of manufacture Textil Oeste Ltda located in Mondai (SC) for three machines. The results based on discriminant scores is more accurate, that the MAPE and MSE.

Decomposing The Fraction Of Missing Information Into Auxiliary Variable Contributions For Monitoring Survey Data

✦Rebecca Andridge, The Ohio State University, B-116 Starling-Loving Hall, 320 W. 10th Ave, Columbus, OH 43210, *randridge@cph.osu.edu*

Key Words: Nonresponse, Auxiliary variable, Imputation

The fraction of missing information (FMI) has recently been proposed as an alternative to the response rate for monitoring the quality of survey data. In order for the FMI to inform data collection decisions (e.g., in adaptive designs or for future survey waves), it must be decomposed into the relative contributions of individual auxiliary variables. We investigate these relative contributions in two ways. First, under some simplifying assumptions, we analytically decompose FMI to show the impact of each covariate. Secondly, we discuss several methods for estimating these relative contributions to FMI in practice, and contrast with alternate quality indicators. The methods are illustrated using data from the National Health Interview Survey (NHIS).

Comparison Of Sampling Methods For A School-Based Population

✦Alana Christie, Dept of Biostatistics and Epidemiology, College of Public Health, OUHSC, 801 NE 13th St., CHB-309, Oklahoma City, OK 73104, *alana-christie@ouhsc.edu*; Barbara Neas, Dept of Biostatistics and Epidemiology, College of Public Health, OUHSC

Key Words: systematic, weighted, school-based, response rate, simulation

Since 2002, the Oklahoma Dental Health Services have surveyed the dental health of the third grade population to assess the prevalence of caries and sealants. To compare different sampling methods, we simulate a population based on estimates from the previous 6 years of the Oral Health Needs Assessment and the current school percentage free and reduced lunch (FRL) report. Sampling methods include regional random sampling (current method), regional systematic sampling and state-wide systematic sampling. All samples will be of 36 schools with varying school sizes and selected after randomization by FRL percentage. For the two regional methods, the state is divided geographically into 6 regions and 6 schools selected from each region. Preliminary data indicate small differences from the population values when the complete sample is available. When the response rate decreases, both the estimates and variability are generally larger for the weighted estimates. Additional comparisons will focus on variations in the response rates and estimates obtained with and without weighting. Non-response adjustment will be explored.

The Non-Verbal Response Card Method For Soliciting Responses To Sensitive Questions

✦David P Lindstrom, Population Studies and Training Center, Brown University, Providence, RI 02912, *David_Lindstrom@brown.edu*; Megan Klein Hattori, Population Studies and Training Center, Brown University

Key Words: Survey Response Bias, Sensitive Questions

This study presents results from a new non-verbal response card method for obtaining more accurate responses to questions about sexual knowledge, attitudes and behavior in the context of interviewer-administered questionnaires. The effectiveness of the cards was tested in two random samples of young people ages 13-24 in southwest Ethiopia (n=201 and n=1,269) using a randomized control trial design in which one-half of the sample used the response cards and the other half of the sample provided verbal responses. The non-verbal response card method produces estimates of the prevalence of pre-marital and extra-marital sexual intercourse that are around twice as high as the estimates provided by the conventional verbal response method. We also found that estimates of the percentage of youth who knew where to obtain condoms were approximately 22 percent lower among youth who used the more private and confidential card method as compared to the verbal response method. Results from two rounds of a longitudinal survey of youth in the same area provide additional evidence of the internal reliability of the card method. The non-verbal response card provides a more private and confidential me

428 Contributed Oral Poster Presentations: WNAR

WNAR

Tuesday, August 2, 2:00 p.m.–3:50 p.m.

Multipim: An R Package For Variable Importance Analysis

✦Stephan Johannes Ritter, UC Berkeley, Group in Biostatistics, , *sritter@berkeley.edu*; Nicholas P Jewell, University of California, Berkeley; Alan Hubbard, University of California, Berkeley

Key Words: R packages, variable importance, super learner

We have written an R package, multiPIM, which performs variable importance analysis. The user must input one or more exposures, one or more outcomes, and optionally, one or more covariates to include in the adjustment set. An effect measure (and an associated standard error) is calculated for each exposure-outcome pair. PIM stands for Population Intervention Model. The parameter of interest is a type of attributable risk, and the default is to use a double-robust inverse probability of censoring weighted estimator for this parameter. The default method of estimating the nuisance parameters of each model is to combine several regression algorithms in a super learner. A simulation and a data re-analysis will also be described.

429 Wald Lecture: Random Walks from Statistical Physics (I) ●

IMS, International Chinese Statistical Association, SSC, Committee of Representatives to AAAS

Tuesday, August 2, 4:00 p.m.–5:50 p.m.

Random Walks: Simple and Self-Avoiding

✦Gregory F Lawler, University of Chicago, , *lawler@math.uchicago.edu*

This will be a survey of what is known about a number of models that arise in statistical physics with an emphasis on: simple random walk (including intersection properties), self-avoiding random walk, and loop-erased random walk. I will discuss how the behavior depends on the dimension of space, summarize what is known rigorously, and list some of the difficult open problems. This talk is intended for a general(statistically literate) audience.

430 Deming Lecture

ASA, ENAR, IMS, International Chinese Statistical Association, International Indian Statistical Association, SSC, WNAR

Tuesday, August 2, 4:00 p.m.–5:50 p.m.

The World Is Calling; Should We Answer?

✦Roger Hoerl, GE Global Research, Applied Statistics Laboratory, 1 Research Circle, Niskayuna, NY 12309 USA, *roger.hoerl@ge.com*

A frequently overlooked aspect of Ed Deming's career was that he focused it not on statistical methods per se, but rather on how he could best contribute to society. He didn't let the fact that he was a statistician - a top notch statistician in fact - limit his ability to make a profound impact on the world. In the spirit of Deming, it is obvious that the world is facing some unique challenges today, which at first glance appear to be insurmountable problems. Wars and arms proliferation, rampant corruption, extreme poverty, communicable disease, and malnutrition come immediately to mind. Providing broader access to healthcare and education, while reducing their overall costs, as well as job creation, are also of keen interest globally. No sector of society appears to be providing sufficient leadership to address such issues. The world is calling for help, but no one seems to be answering. Could statisticians answer the call? Do we have the ability to address not only narrow technical questions, but also the big, complex challenges of the 21st century? I believe we do, but not by focusing on perfecting individual tools. We could contribute to society in a much more significant way by thinking bigger, and creatively integrating statistical and non-statistical tools to develop innovative approaches to attack the large, complex, unstructured problems the world is facing. Ron Snee and I refer to such integration of multiple statistical methods into overall strategies to attack complex, unstructured problems as statistical engineering. As one example of how we can utilize our unique skills and training to help address society's greatest needs, I will share some of my own experiences in the fight against the HIV/AIDS pandemic.

431 ASA Presidential Address and Awards

ASA, ENAR, IMS, International Chinese Statistical Association, International Indian Statistical Association, WNAR

Tuesday, August 2, 8:00 p.m.–9:30 p.m.

Statistics: An All-Encompassing Discipline

✦Nancy L. Geller, Office of Biostat Research, 2 Rockledge Centre Suite 9093, Bethesda, MD 20892-7913, *nancylgeller@gmail.com*

Presentation of Awards

✦Sastry Pantula, NSF, , *pantula@stat.ncsu.edu*