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

Section on Bayesian Statistical Science

Monday, August 1, 7:00 a.m.–8:15 a.m.

21St Century Computational Statistics: Clouds, Gpus And What'S Next?

✦Marc Suchard, UCLA, 695 Charles E. Young Dr., South, Los Angeles, CA 90095, *msuchard@ucla.edu*

Key Words: computational statistics, parallelization, optimization, MCMC, Bayesian modeling

Data-rich, 21st century science and commerce is challenging Computational Statistics. This round-table discusses how distributed cloud computing, multi-core and massive many-core parallelization through graphics processing units (GPUs) can provide several orders-of-magnitude performance boosts in simulation, optimization and stochastic search methods for fitting and evaluating complex models with increasingly large data. Software libraries, hardware technology and novel statistical algorithms are all on the table.

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

Section on Physical and Engineering Sciences

Monday, August 1, 7:00 a.m.–8:15 a.m.

New Directions and Methods in Process Monitoring

✦William H. Woodall, Virginia Tech, Department of Statistics, Blacksburg, VA 24061-0439, *bwoodall@vt.edu*

Key Words: control chart, statistical process control, multivariate quality control, CUSUM charts, image data

We will discuss some of the latest developments in process monitoring and statistical process control. The topics will depend largely on the interests of the participants, but can include the monitoring of image data, profile monitoring, risk-adjusted monitoring in healthcare, and the monitoring of "high quality" attribute processes.

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

Section on Quality and Productivity

Monday, August 1, 7:00 a.m.–8:15 a.m.

Bayesian Reliability

✦Alyson Wilson, Iowa State University, 2105 Snedecor Hall, Department of Statistics, Ames, IA 50011, *agw@iastate.edu*

Key Words: Bayesian, reliability, MCMC, assurance test

With the advent of improved computational tools (especially Markov chain Monte Carlo), Bayesian methods are becoming more widely used in reliability. This roundtable will focus on practical issue related to Bayesian reliability assessments, including prior distributions, computing, using covariates, designing assurance tests, communicating results, and experiences with applications.

93 Section on Statistical Education A.M. Roundtable Discussion (fee event)

Section on Statistical Education

Monday, August 1, 7:00 a.m.–8:15 a.m.

Becoming a Teacher of Statistics

✦Michelle Everson, University of Minnesota, Dept. of Educational Psychology, 250 Ed Sci Bldg, 56 East River Road, Minneapolis, MN 55455, *gaddy001@umn.edu*

Key Words: introductory statistics, teaching, GAISE, statistics education

Are you just finishing a Ph.D. program and preparing for your first teaching job? Or, are you about to start that first teaching job this coming fall? This roundtable will focus on discussion of current recommendations and best practices (GAISE) for teaching introductory statistics at the college level. In addition, many teaching resources will be shared, and discussion will focus on addressing any questions or concerns new teachers have when it comes to surviving that first teaching experience. Come and commiserate with others who are also new to teaching statistics, and learn more about becoming a part of the growing statistics education community!

Quality Assurance in Online Courses: How Do We Establish It?

✦Sue Schou, Idaho State University, PO Box 4043, Pocatello, ID 83205, *schosue@isu.edu*

Key Words: online, quality assurance, statistics education

Participate in a lively discussion about ensuring online courses deliver a similar quality product as your traditionally taught courses. Some of the questions we will address are: 1) How do we evaluate the effectiveness of online courses? 2) Can instructor evaluation for online courses be conducted in the same manner as traditionally taught courses? 3) How do we determine if an online course is adequately prepared for the online learning environment? and 4) How do we evaluate publisher online course products? These questions will serve only as the starting point!

94 Section on Statistics in Defense and National Security A.M. Roundtable Discussion (fee event)

Section on Statistics in Defense and National Security
Monday, August 1, 7:00 a.m.–8:15 a.m.

Supporting National Security Policy With Science And Technology

✦Sallie Keller, IDA Science and Technology Policy Institute, 1899 Pennsylvania Avenue, NW, Suite 520, Washington, DC 20006, *skeller@ida.org*

Key Words: national security, science, technology

Science, technology, and policy --- where the rubber meets the road for national security. We will discuss the nuances of how innovations in science and technology can best support national security policy, and how national security policy needs to be modified as a result of innovations in science and technology.

95 Section on Statistics in Epidemiology A.M. Roundtable Discussion (fee event)

Section on Statistics in Epidemiology
Monday, August 1, 7:00 a.m.–8:15 a.m.

Gerontologic Biostatistics: Time To Consider Best Practices?

✦Terrence E. Murphy, Yale University School of Medicine, 300 George St, Suite 775, New Haven, CT 06511, *terrence.murphy@yale.edu*

Key Words: gerontologic biostatistics, , non-informative censoring, competing events, joint models, best practices

The subdiscipline known as gerontologic biostatistics is centered around a systematic and comprehensive treatment of the special statistical challenges that arise when conducting research with older persons. These statistical challenges include the design and analysis of multicomponent interventions, the simultaneous modeling of multiple outcomes, and consideration for the timing and occurrence of the comorbidities that inevitably accompany the aging process of older persons. In this year's roundtable we pose the idea of whether it would be helpful to start categorizing different analytical approaches into a system of "best practices", a practice commonly used in industry. For purposes of discussion, we will examine the hierarchy of analytical preferences when performing longitudinal modeling of gerontological outcomes. These include simple survival modeling that assumes non-informative censoring, the use of competing risk models, and the application of joint modeling techniques to account for the predictable loss of followup due to death. Attendees will be asked to name an area of analysis within the subdiscipline where the establishment of "best practices" might be useful.

96 Section on Teaching of Statistics in the Health Sciences A.M. Roundtable Discussion (fee event)

Section on Teaching of Statistics in the Health Sciences
Monday, August 1, 7:00 a.m.–8:15 a.m.

Quick One-Page Tutorials Promoting Statistical Literacy

✦Becki Bucher Bartelson, Rocky Mountain Poison and Drug Center, 777 Bannock Street, Denver, CO 80204, *Becki.Bucher-Bartelson@rmpdc.org*; Lynn Ackerson, Kaiser Permanente Division of Research; Amanda Allshouse, Colorado School of Public Health; Samantha MaWhinney, Colorado School of Public Health

Key Words: Biostatistical Literacy, Education, Health Sciences

Health science researchers often seek statistical guidance outside of normal business hours or request focused supplementary reading materials on particular statistical issues. To meet these needs we have developed a series of one page handouts on a variety of commonly used statistical methods/topics such as descriptive statistics, hypothesis tests, confidence intervals, t-tests, contingency tables and regression. Tutorials were designed with an emphasis on statistical literacy versus the "how-to" approach. At this round table we will discuss tutorial goals, their development, target audience and dissemination strategies. Lastly, we will present the results of a small focus group evaluation of the tutorials by a group of toxicology fellows. We will encourage an active group discussion of these topics and thoughts regarding the role of quick tutorials in promoting statistical literacy in the health sciences. The goal of this roundtable is to engage in an exchange of ideas that will be beneficial to all present.

97 Introductory Overview Lecture: Statistics and Evidence-Based Medicine

ASA
Monday, August 1, 8:30 a.m.–10:20 a.m.

Statistics and Evidence-Based Medicine

✦Christopher Schmid, Tufts Medical Center, 800 Washington St, Box 63, Boston, MA 02111, *cschmid@tuftsmedicalcenter.org*

Evidence-based medicine (EBM) is the use of the best available evidence gathered using scientific methods to make decisions about the care of patients. Statistics plays a key role in EBM. Sources of evidence for the risks and benefits of interventions and tests include clinical trials, observational studies, diagnostic tests, meta-analysis, decision and cost effectiveness analysis. Many clinical decisions relate to multivariate choices that focus on competing outcomes, treatments, populations and time frames. Modern EBM and comparative effectiveness research (CER) use multivariate statistical methods to synthesize existing evidence and suggest areas of most benefit for future research. This talk will review statistical tools used in EBM emphasizing multivariate techniques and the contributions that statistics can make in efficiently generating, synthesizing and interpreting evidence applicable both to population and individual level clinical care

98 Real-Life Ethical Dilemmas Encountered in the Practice of Statistics: They Happen Everywhere ■●

Committee on Professional Ethics, Section on Statistical Education, Section on Statistical Consulting

Monday, August 1, 8:30 a.m.–10:20 a.m.

Bush v. Gore in Florida 2000: A Questionable Margin of Victory and a Statistician's Dilemma in Testifying from 11th-Hour Data

✦Arlene Ash, University of Massachusetts Medical School, Worcester, MA , *Arlene.Ash@umassmed.edu*

Key Words: ethics, elections, testimony, consulting

Dr. Ash agreed to testify in December 2000 regarding the effect on the US presidential election of about 1200 imperfectly-filled-out and questionably re-enfranchised absentee ballots originally solicited by the Republican Party of Martin County, Florida. The night before she was due to testify, some newly available data called into question the applicability to Martin County of some of the original (statewide polling) data on which she had relied in making preliminary estimates of the effect of what now looked like approximately 900 ballots on the Bush v. Gore margin of victory (then 537). The defect was fairly subtle and it seemed likely that no one else would notice, but could Dr. Ash honorably move forward, and if so how?

An Ethical Issue Related to Teaching at the Undergraduate Level

✦Katherine Taylor Halvorsen, Smith College, Clark Science Center, 44 College Lane, Northampton, MA 01063, *khalvors@smith.edu*

Key Words: Multiple testing, Data dredging, Ethics, Consulting

Students in my introductory statistics course for undergraduate math and science majors complete a term project requiring them to design a study, collect and analyze data, and write a report. The few graduate students who enroll usually use their masters-thesis data as the basis for their project. After taking my fall course, one graduate student returned frequently during the spring for consulting on other analyses for her thesis. We discussed ways to test her hypotheses and she carried out the tests. When she showed me her final results, I noted that she had used the t-test to do data dredging, looking for any possible associations she could find. In her summary report, she included tables showing only the significant tests, declaring the tests significant without using multiple comparison procedures to adjust for the multiple testing. I explained that reporting the tests significant without adjusting for multiple comparisons misleads readers. She said her thesis advisors were pleased with her work and were expecting her to publish a paper from her thesis, including the tables she showed me. All she wanted from me was wording to explain the results.

Ethical Dilemmas in Writing and Publishing Statistical Texts

✦Jeffrey Witmer, Oberlin College, Mathematics Department, 10 N. Professor St, Oberlin, OH 44074, *jeff.witmer@oberlin.edu*

Key Words: textbook, ethics, contract

Writing a textbook takes a lot of work, but a successful product brings in royalties. Is it ethical to collect royalty income by adopting a text that you've written as the required text in a course that you are teaching? And what about getting a contract with a publisher in the first place? How can you form a good working relationship with a publisher preserving your rights? I will share opinions and experiences.

Ethical Dilemmas of an Isolated Statistician

✦Donald Bentley, Pomona College, 1826 Roanoke Rd, Claremont, CA 91711, *dbentley@pomona.edu*

Key Words: ethical dilemmas, liberal arts college, biblical scholarship, consulting

This paper presents two ethical dilemmas faced by an "isolated statistician" teaching in a small liberal arts college. The first scenario arose while consulting for a start-up pharmaceutical firm, and dealt with a question about the integrity of data provided by a clinical investigator. The second dealt with an issue in unusual areas for the application of statistics, biblical scholarship and archaeology, in which the statistician was uniquely qualified. The decision to participate in this project created the initial dilemma. After describing the scenarios and identifying the dilemmas faced by the applied statistician, the actions taken and ultimate consequences will be provided. Discussion will include problems which arise due to the "isolated statistician's" environment.

99 Showcasing recent papers from the Journal of Agriculture, Biological and Environmental Statistics ■●

ENAR, Section on Statistics and the Environment, WNAR

Monday, August 1, 8:30 a.m.–10:20 a.m.

Air Pollution and Pre-Term Pregnancy: Identifying Critical Windows of Exposure

✦Montse Fuentes, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695, *montse_fuentes@ncsu.edu*; Brian Reich, North Carolina State University

Key Words: multivariate statistics, spatial statistics, particulate matter, ozone, pregnancy outcomes, probit spatial models

A major methodological challenge in the study of environmental exposures and birth outcomes is the identification of a critical window of exposure. The developing organ systems of the fetus may be more vulnerable to exposures to environmental toxicants during these critical windows because of a variety of factors, including higher rates of cell proliferation or changes in metabolism. We introduce new Bayesian spatial shrinkage methods to control potential multicollinearity of exposure due to multi-pollutants exposure and multiple lags. This framework facilitates the search for susceptible periods of exposure during fetal development, and sheds additional light on exposures previously examined in isolation or under strict assumptions about the nature of the association. We apply our methods to geo-coded birth outcome data from the state of Texas (1997-2004) to identify the critical windows of the pregnancy where increased exposure to fine particulate

matter and ozone is particularly harmful. Our results indicate the susceptible window for higher preterm probabilities is mid-first trimester for the fine PM and beginning of the first trimester for the ozone.

Analyzing Spatial Directional Data with Measurement Error Using Wrapped Gaussian Processes

✦Alan E Gelfand, Department of Statistical Science, Duke University, Durham, NC 27708-0251, *alan@stat.duke.edu*

Key Words: Bayesian kriging, concentration parameter, Gaussian process, hierarchical model, latent variables, resultant length

Circular data arise in oceanography (wave directions) and meteorology (wind directions), and, more generally, with measurements recorded in degrees or angles on a circle. In this talk we introduce a fully model-based approach to handle circular data in the case of measurements taken at spatial locations, anticipating structured dependence between these measurements. We formulate a wrapped Gaussian spatial process model for this setting, induced from a customary inline Gaussian process. We build a hierarchical model to handle this situation and show how to fit this model straightforwardly using Markov chain Monte Carlo methods. Our approach enables spatial interpolation and can accommodate measurement error. We illustrate with a set of wave direction data from the Adriatic coast of Italy, generated through a complex computer model.

A Bayesian Functional Data Model for Predicting Forest Variables Using High-Dimensional Waveform LiDAR Over Large Geographic Domains

✦Andrew Oliver Finley, Michigan State University, Natural Resources Building, Michigan State Univers, East Lansing, MI 48824, *finleya@msu.edu*; Sudipto Banerjee, University of Minnesota; Bruce Cook, NASA's Goddard Space Flight Center

Key Words: MCMC, predictive process, spatial GLM, spatial process, Hierarchical model

Recent advances in remote sensing, specifically waveform Light Detection and Ranging (LiDAR) sensors, provide the data needed to quantify forest variables at a fine spatial resolution over large domains. We define a framework to couple a spatial latent factor model with forest variables using a fully Bayesian functional spatial data analysis. Our proposed modeling framework explicitly: 1) reduces the dimensionality of signals in an optimal way (i.e., preserves the information that describes the maximum variability in response variable); 2) propagates uncertainty in data and parameters through to prediction, and; 3) acknowledges and leverages spatial dependence among the regressors and model residuals to meet statistical assumptions and improve prediction. The dimensionality of the problem is further reduced by replacing each factor's Gaussian spatial process with a reduced rank predictive process. The proposed modeling framework is illustrated using waveform LiDAR and spatially coinciding forest inventory data collected on the Penobscot Experimental Forest, Maine.

Spatio-Temporal Models for Oceanic Data

✦Bruno Sanso, University of California, Santa Cruz, Department of Applied Mathematics and Statistics, School of Engineering, Santa Cruz, 95064 Canada, *bruno@ams.ucsc.edu*; Ricardo Lemos, NOAA/NMFS Environmental Research Division

Key Words: Bayesian Hierarchical Models, Spatio-temporal models, Climatology

We present a review of Bayesian hierarchical models for the reconstruction of oceanic properties. By using a hierarchical spatio-temporal model we are able to consider long series of observations irregularly scattered in space and time. Additionally, we account for observational errors and incorporate structural information about the underlying physical processes. Our latest development is HOMER: a Hierarchical Ocean Model for Extended Reconstructions. Its goal is to obtain smooth three dimensional fields of temperature and salinity, as well as long term climatologies, on a monthly time scale. We develop carefully designed Markov chain Monte Carlo algorithms on distributed machines to handle massive datasets that correspond to long time series and large geographical domains.

100 Bayesian Methods for Causal Inference ■●

Section on Bayesian Statistical Science, Section on Health Policy Statistics, Section on Statistics in Epidemiology

Monday, August 1, 8:30 a.m.–10:20 a.m.

Bayesian Methods for Causal Inference on Mediators with Application to Weight Management Clinical Trials

✦Michael Daniels, University of Florida, Department of Statistics, 102C Griffin-Floyd Hall, Gainesville, FL 32611 USA, *mdaniels@stat.ufl.edu*; Jason Roy, University of Pennsylvania; Chanmin Kim, University of Florida; Joseph Hogan, Brown University; Michael Perri, University of Florida

Key Words: Mediation, Sensitivity analysis

We propose a nonparametric Bayesian approach to estimate the natural direct and indirect effects of a mediator in the setting of a continuous mediator and a binary response. Several conditional independence assumptions are introduced (with corresponding sensitivity parameters) to make these effects identifiable from the observed data. This approach is used to assess mediation in a recent weight management clinical trial.

Bayesian Inference in Partially Identified Models

✦Paul Gustafson, University of British Columbia, Department of Statistics, 333-6356 Agricultural Rd., Vancouver, BC V6T1Z2 Canada, *gustaf@stat.ubc.ca*

Key Words: Bayesian inference, causal inference, identification

Identification can be a major issue in causal modeling contexts, and in contexts where observational studies have various limitations. Sometimes partial identification arises. In Bayesian terms, this implies that as the sample size grows, the support of the posterior distribution on

the target parameter converges to a set which is smaller than the support of the prior distribution but larger than a single point. We discuss properties of Bayesian inference in partially identified models, with examples drawn from causal modeling contexts. Special attention is paid to the performance of posterior credible intervals arising from partially identified models.

Causal Inference Using Bayesian Nonparametric Modeling

✦Jennifer Hill, New York University, 10012 US, *jennifer.hill@nyu.edu*

Key Words: causal inference, Bayesian, propensity score, missing data, nonparametric, common support

Researchers have long struggled to identify causal effects in non-experimental settings. Many recently proposed strategies assume ignorability of the treatment assignment mechanism and require fitting two models-one for the assignment mechanism and one for the response surface. An alternate strategy is proposed that instead focuses on very flexibly modeling just the response surface using a Bayesian nonparametric modeling procedure, Bayesian Additive Regression Trees (BART). BART has advantages with regard to ease of use, ability to handle a large number of predictors, coherent uncertainty intervals, and natural accommodation of continuous treatment variables and missing data for the outcome variable. BART also naturally identifies heterogeneous treatment effects. BART has been shown to produce more accurate estimates of average treatment effects compared competitors in the propensity score world in the nonlinear simulation situations examined. Further, it is highly competitive in linear settings with the "correct" model, linear regression. The approach can also be used to identify areas that lack common support. Extensions that allow for covariate missing data will be discussed.

101 The Future of Statistical Computing Environments ■

Section on Statistical Computing, Section on Statistical Graphics
Monday, August 1, 8:30 a.m.–10:20 a.m.

Some Developments for the R Engine

✦Luke Tierney, University of Iowa, Department of Statistics & Actuarial Science, 241 Schaeffer Hall, Iowa City, IA 52242, *luke@stat.uiowa.edu*

Key Words: computing environment, compilation, parallel computing

The R language for statistical computing and graphics has become a major framework for both statistical practice and research. This talk will describe some current efforts on improvements to the core computational engine, including work on compilation of R code and efforts to take advantage of multiple processor cores.

The Q Project: Explorations Using the Parrot Virtual Machine

✦Michael Kane, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511 USA, *michael.kane@yale.edu*; John Emerson, Yale University

Key Words: Parrot, virtual machine, statistical software, R, S, bytecode

Parrot (http://www.parrot.org) is a virtual machine designed to efficiently compile and execute bytecode for dynamic languages. As of late 2010, it hosts a variety of language implementations in various stages of development, including Tcl, Javascript, Ruby, Lua, Scheme, PHP, Python, Perl 6, APL, and a .NET bytecode translator. It is designed to provide interoperability between languages that compile to it. This talk reports on preliminary attempts to implement the basics of the S language in Parrot, a project code-named "The Q Project."

Taking Statistical Computing Beyond S and R

✦Simon Urbanek, AT&T Labs - Research, 180 Park Avenue, Florham Park, NJ 07932, *urbanek@research.att.com*

Key Words: R, statistical computing, object system

Aleph is an open-source project to create the next generation of statistical computing software, possibly as a successor to R. The goal is to provide a modern, flexible system suitable for statistical analysis. All aspects of the project are currently experimental and up for discussion. The current experimental implementation is written in C and features its own C-level object system. Although other approaches were considered, so far we have decided to use the core S/R language as the basis for the Aleph language. The syntax should be very similar such that it should be possible to translate packages from R to Aleph if necessary (not really a requirement but a side-effect). However, some aspects of the language (lazy evaluation, pass-by-value, etc.) are not mandatory. One of the key points is that Aleph has its own object system. Some major features are first-class class objects, fundamentally complete class structure, type matched function arguments.

CXXR: An Ideas Hatchery for Future R Development

✦Andrew Runnalls, University of Kent, UK, School of Computing, University of Kent, Canterbury, CT2 7NF UK, *A.R.Runnalls@kent.ac.uk*

Key Words: R, C++, refactoring, computing, object-oriented

The continued growth of CRAN is testament to the increasing number of developers engaged in R development. But far fewer researchers have experimented with the R interpreter itself. The code of the interpreter, written for the most part in C, is structured in a way that will be foreign to students brought up with object-oriented programming, and the available documentation, though giving a general understanding of how the interpreter works, does not really enable a newcomer to start modifying the code with any confidence. The CXXR project is progressively refactoring the interpreter into C++, whilst all the time preserving existing functionality. By restructuring the code into tightly encapsulated and carefully documented classes, CXXR aims to open up the interpreter to more ready experimentation by statistical computing researchers. This paper focusses on two example tasks:

(a) adding the capability to track the provenance of R objects, and (b) providing, as a package, a new type of data vector. The paper shows how CXXR greatly facilitates these tasks by internal changes to the structure of the interpreter, and by offering a higher-level interface for packages to exploit.

102 Recent Advances in Modeling Spatial and Temporal Data ●

Section on Statistics and the Environment, International Indian Statistical Association, Section on Physical and Engineering Sciences, Section on Statistical Computing, Section on Statistics in Epidemiology

Monday, August 1, 8:30 a.m.–10:20 a.m.

Non-Gaussian Spatiotemporal Modelling Through Scale Mixing

✦Thais C O Fonseca, Unirio, Departamento de Estatistica - Unirio, CCET - Avenida Pasteur, 458 - Urca, Rio de Janeiro, 22290-240 Brazil, *thais.fonseca@uniriotec.br*; Mark F J Steel, Warwick University

Key Words: non-Gaussian processes, outlier detection, Bayesian inference, Mixture models, Nonseparable covariances

We construct non-Gaussian processes that vary continuously in space and time with nonseparable covariance functions. Stochastic modelling of phenomena over space and time is important in many areas of application. We start from a general and flexible way of constructing valid nonseparable covariance functions through mixing over separable covariance functions. We then generalize the resulting models by allowing for individual outliers as well as regions with larger variances. We induce this through scale mixing with separate positive-valued processes. Smooth mixing processes are applied to the underlying correlated processes in space and in time, thus leading to regions in space and time of increased spread. We also apply a separate uncorrelated mixing process to the nugget effect to generate individual outliers. We consider posterior and predictive Bayesian inference with these models and implement this through a Markov chain Monte Carlo sampler. Finally, this modelling approach is applied to temperature data in the Basque country.

An Approach to Modeling Asymmetric Multivariate Spatial Covariance Structures

✦Bo Li, Purdue University, 150 North University St., Department of Statistics, Purdue University, West Lafayette, IN 47907, *boli@purdue.edu*; Heping Zhang, Yale University

Key Words: Asymmetry, Bivariate Matern, Intrinsic model, Multivariate covariance function, Symmetry

We propose a framework in light of the delay effect to model the asymmetry of multivariate covariance functions that are often exhibited in real data. This general approach can endow any valid symmetric multivariate covariance function with the ability of modeling asymmetry and is very easy to implement. Our simulations and real data examples show that asymmetric multivariate covariance functions based on our approach can achieve remarkable improvements in prediction over the symmetric model.

A Valid Matern Class of Cross-Covariance Functions for Multivariate Random Fields with Any Number of Components

✦Tatiyana V. Apanasovich, Thomas Jefferson University, 1015 Chestnut str M100, Philadelphia, PA 19107, *tatiyana.apanasovich@jefferson.edu*; Marc G. Genton, Texas A&M University; Ying Sun, Texas A&M University

Key Words: Cokriging, Multivariate, Valid cross-covariance, Spatial

We introduce a valid parametric family of cross-covariance functions for multivariate spatial random fields where each component has a covariance function from a well-celebrated Matern class. Unlike previous attempts, our model indeed allows for various smoothnesses and rates of correlation decay for any number of vector components. We present the conditions on the parameter space that result in valid models with varying degrees of complexity. Practical implementations, including reparametrizations to reflect the conditions on the parameter space and an iterative algorithm to increase the computational efficiency, are discussed. We perform various Monte Carlo simulation experiments to explore the performances of our approach in terms of estimation and cokriging. The application of the proposed multivariate Matern model is illustrated on two meteorological datasets: Temperature/Pressure over the Pacific Northwest (bivariate) and Wind/Temperature/Pressure in Oklahoma (trivariate).

Analysis of Massive Data Set Through Compactly Supported Covariance Functions

✦Emilio Porcu, , *eporcu@uni-goettingen.de*

We propose new classes of covariance functions for vector-valued random fields having the additional feature of being compactly supported, which is desirable for practitioners working on massive spatial data sets. In particular, we propose compactly supported matrix-valued mappings generated by the Wu class of covariance functions as well as its generalizations to the so-called Buhmann and Gneiting-Wendland classes. An application to Pacific Ocean temperature and pressure data, as well as a simulation study, illustrate the features of such models.

103 Recent Research on Total Survey Error ■●

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

Monday, August 1, 8:30 a.m.–10:20 a.m.

How Much of Interviewer Variance Is Really Nonresponse Error Variance?

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

Key Words: Interviewer Variance, Nonresponse Error, Measurement Error, Total Survey Error, Interviewer Effects, CATI

Intra-interviewer correlation can arise when answers from survey respondents interviewed by the same interviewer are more similar to each other than answers from other respondents, decreasing the precision of survey estimates. Estimation of this parameter in practice, however, only uses respondent data. The potential contribution of variance in nonresponse errors between interviewers to this correlation has been largely ignored: responses within interviewers may appear correlated because the interviewers successfully obtain cooperation from different pools of respondents. This study attempts to fill this gap by analyzing a unique survey data set, which includes both true values and reported values for respondents and arises from a CATI sample assignment that approximates interpenetrated assignment of subsamples to interviewers. This data set enables the decomposition of interviewer variance in means of respondent reports into nonresponse error variance and measurement error variance across interviewers. We show that in cases where there is substantial interviewer variance in reported values, the interviewer variance may arise from nonresponse error variance across interviewers.

Nonresponse and Measurement Error: Relationship, Relative Magnitude, and Correction

✦Andy Peytchev, RTI, 3040 Cornwallis Rd, Research Triangle Park, NC 27709, *apeytchev@rti.org*

Key Words: Nonresponse Bias, Measurement Error, Multiple Imputation, MSE, Total Survey Error, Paradata

Commonly employed weighting methods to address nonresponse generally lead to reduced precision, spurring tradeoffs between bias and variance. No corrections for measurement error are commonly incorporated, while variances of survey estimates are typically penalized when ample auxiliary information is available to correct for nonresponse. Multiple imputation can be used for unit nonresponse, item nonresponse, as well as measurement error, making use of frame data, paradata, and survey data with varying levels of missingness. Unlike nonresponse weighting adjustments, it can reduce variances, thus also reducing mean squared error of estimates (MSE). This approach was applied to data from NSFG cycle 5, which includes a rich sampling frame, paradata, and replicate measures less prone to measurement error. Based on multiply-imputed data, measurement error bias was almost three times larger than nonresponse bias. Although the same conclusion was reached through weighting, multiple imputation yielded substantially lower variance estimates and estimates of MSE.

Challenges in Minimizing Nonsampling Errors in an International Assessment of Adult Competencies

✦Leyla K. Mohadjer, Westat, 1600 Reserach B, Rockville, MD 20850, *leylamohadjer@westat.com*

Key Words: Nonsampling errors, international surveys, standards and guidelines, quality control

The Programme for the International Assessment of Adult Competencies (PIAAC) is a multi-cycle international survey of assessment of adult skills and competencies sponsored by the Organization for Economic Cooperation and Development (OECD). PIAAC will collect background information and administer an assessment of cognitive skills to measure participants' general levels of literacy and numeracy. In PIAAC, as in any survey, it is a challenge to minimize potential survey errors, which may be due to such factors as the sample design or selection, the measurement instruments, data collection or processing problems, weighting and estimation difficulties, etc. Furthermore, in a multi-national survey such as PIAAC, there are challenges associated with the diversity of cultures, languages and other practices among countries. No single survey design will be effective for every country. Nevertheless, because of survey complexities and the possibility of different practices among countries, it is important to standardize the PIAAC survey procedures as much as practically possible. This paper presents PIAAC's standards and guidelines, and the preliminary results gathered through a field test.

Repercussions of Nonresponse Follow-Up for Measurement Error

✦Frauke Kreuter, University of Maryland, 1218 LeFrak Hall, College Park, MD 20742, *fkreuter@survey.umd.edu*

Surveys face the threat of nonresponse bias if respondents differ systematically from nonrespondents. As safeguard survey researchers try to maintain high response rates through increased recruitment efforts. However, such efforts are costly and can prevent timeliness of statistic production, while they may not reduce nonresponse bias. In addition pressing sample cases to respond can backfire and compromise measurement quality, a hypothesis that has been put forward repeatedly in the past. This presentation will review investigations of measurement error in nonresponse follow-up studies, and will among others present results form a record linkage study that jointly examined nonresponse and measurement error. This particular study revealed an interesting case of counteracting effects of nonresponse and measurement error. Here mean square error increased despite a reduction in nonresponse bias and little to no increase in measurement error. General implications of such counteracting effects will be discussed.

104 Analysis of Genome-wide Association Studies: methods from the CHARGE consortium of cohort studies ■

WNAR, ENAR

Monday, August 1, 8:30 a.m.–10:20 a.m.

A Multivariate Approach on Genome-Wide Association Studies (GWAS) by Modeling Multiple Traits Simultaneously to Identify Pleiotropic Genetic Effects

✦Yi-Hsiang Hsu, Hebrew SeniorLife Institute for Aging Research and Harvard Medical School, 1200 Centre Street, Research, Boston, MA 02131, *yihsianghsu@hsl.harvard.edu*; Xing Chen, Harvard School of Public Health; David Karasik, Hebrew Sneiorlife Institute for Aging Research and Harvard Medical School; Kathryn Lunetta, School of Public Health, Boston University; Douglas Kiel, Hebrew SeniorLife Institute for Aging Research and Harvard Medical School

Key Words: GWAS, PLEIOTROPY, SNP, MUTIVARIATE, CORRELATED PHENOTYPES

Pleiotropy occur when multiple traits were affected independently by same genetic variants. Due to correlation among traits and moderate genetic effects of GWAS, it is inefficient to detect pleiotropy by univariate analytical framework. We propose here a new approach to test pleiotropy on GWAS using a two-stage strategy: in the first stage, we performed a multi-phenotype GWAS by modeling traits simultaneously using our newly developed empirical-weighted linear-combined test statistics (eLC); and then, we tested the pleiotropy using a simplified structure-equation-modeling on selected SNPs from the first stage. eLC directly combines correlated test-statistics to an overall association test with a weighted sum of univariate statistics to maximize the information obtained from each univariate analysis. Using GWA16 simulated dataset, our eLC approach has outperformed the simple look-up on the overlaps among univariate GWAS and other multivariate methods (such as MANOVA, GEE and PCA). We applied our approach to data from the CHARGE and GEFOS consortia and identified pleiotropic genetic effects on reproductive phenotypes (age at menarche and age at menopause) and bone mineral density.

Gene-Based Tests of Association

✦Joel S Bader, Johns Hopkins University, 3400 N. Charles St., 201C Clark Hall, Baltimore, MD 21218, *joel.bader@jhu.edu*; Hailiang Huang, Johns Hopkins University; Alvaro Alonso, University of Minnesota; Dan E Arking, Johns Hopkins University

Key Words: gwas, genome-wide association, genetics, bayesian model selection, bioinformatics, genomics

Genome-wide association studies (GWAS) are now used routinely to identify SNPs associated with complex human phenotypes. In several cases, multiple variants within a gene contribute independently to disease risk. Here we introduce a novel Gene-Wide Significance (GWiS) test that uses Bayesian model selection to identify the number of independent effects within a gene, which are combined to generate a stronger statistical signal. Permutation tests provide p-values that correct for the number of independent tests genome-wide and within each genetic locus. When applied to a dataset comprising 2.5 million SNPs in up to 8,000 individuals measured for various electrocardiography (ECG) parameters, this method identifies more validated associations than conventional GWAS approaches. The method also provides, for the first time, a systematic assessment of the fraction of disease-associated genes housing multiple independent effects, observed at 35-50% of loci in our study. This method can be generalized to other study designs, and provides gene-based p-values that are directly compatible for pathway-based meta-analysis.

Analysis of Genome-Wide Association Studies: Methods from the Charge Consortium of Cohort Studies

✦Kenneth M Rice, Department of Biostatistics, University of Washington, F600 HSB, Box 357232, Seattle, WA 98195-7232 USA, *kenrice@u.washington.edu*

Key Words: Genome-wide studies, GxE interaction, stratification

Genome-wide association studies of gene-environment interaction (GxE GWAS) are becoming popular. As with main effects GWAS, quantile-quantile plots (QQ-plots) and Genomic Control are being used to assess and correct for population substructure. However, in GxE work QQ-plots approaches can be seriously misleading, as we illustrate; they may give strong indications of substructure when absolutely none is present. In this talk, we use simulation and theory to show how and why spurious QQ-plot inflation occurs in GxE GWAS, and how this differs from main-effects analyses. We also explain how simple adjustments to standard regression-based methods used in GxE GWAS can alleviate this problem, and describe circumstances under which these 'fixes' will be most practically important. Examples will be drawn from several CHARGE analyses of GxE interactions. This is joint work with Arend Voorman, Thomas Lumley and Barbara McKnight.

105 Recent development in Monte Carlo methods: theory and scientific applications ■●

IMS, International Chinese Statistical Association

Monday, August 1, 8:30 a.m.–10:20 a.m.

Sequential Monte Carlo Methods in Protein Folding

✦Samuel Kou, Harvard University, Department of Statistics, 1 Oxford Street, Cambridge, MA 02138 USA, *kou@stat.harvard.edu*

Key Words: energy function, structure predictoin, conditional sampling, sequential growth, configurational bias Monte Carlo, protein conformation

Predicting the native structure of a protein from its amino acid sequence is a long standing problem. A significant bottleneck of computational prediction is the lack of efficient sampling algorithms to explore of the configuration space of a protein. In this talk we will introduce a sequential Monte Carlo method to address this challenge: fragment regrowth via energy-guided sequential sampling (FRESS). The FRESS algorithm combines statistical learning (namely, learning from the protein data bank) with sequential sampling to guide the computation, resulting in a fast and effective exploration of the configurations. We will illustrate the FRESS algorithm with both lattice protein model and real proteins.

Recent Advances in Optimal Scaling of MCMC Algorithms

✦Natesh S Pillai, Harvard University, 1 Oxford St, Cambridge, MA 02138 USA, *pillai@stat.harvard.edu*

Key Words: optimal scaling, Langevin Diffusions, Hybrid Monte Carlo

Most of the research efforts so far on MCMC were focused on obtaining estimates for the mixing times of the corresponding Markov chain. In this talk wee will discuss optimal scaling of MCMC algorithms in high dimensions where the key idea is to study the properties of the proposal distribution as a function of the dimension. This point of view gives us new insights on the behavior of the algorithm, such as precise

estimates of the number of steps required to explore the target distribution, in stationarity as a function of the dimension of the state space. In the first part of the talk, we will describe the main ideas and discuss recent results on high dimensional target measures arising in the context of statistical inference for mathematical models representing physical phenomena. In the second part of the talk, we will discuss the Hybrid Monte Carlo Algorithm (HMC) and answer a few open questions about its efficiency in high dimensions. We will also briefly discuss applications to parallel tempering, Gibbs samplers and conclude with concrete problems for future work.

Calibrated Path Sampling and Stepwise Bridge Sampling

✦zhiqiang tan, rutgers univ, 08854 USA, *ztan@stat.rutgers.edu*

Key Words: Normalizing constant;, Bridge sampling, Path sampling

Consider probability distributions with unnormalized density functions indexed by parameters on a 2-dimensional grid, and assume that samples are simulated from distributions on a subgrid. Path sampling uses samples along a 1-dimensional path to compute each integral. However, different choices of the path lead to different estimators, which should ideally be identical. We propose calibrated estimators by the method of control variates to exploit such constraints for variance reduction. We also propose biquadratic interpolation to approximate integrals with parameters outside the subgrid, consistently with the calibrated estimators on the subgrid. These methods can be extended to compute differences of expectations through an auxiliary identity for path sampling. Furthermore, we develop stepwise bridge-sampling methods in parallel but complementary to path sampling. In three simulation studies, the proposed methods lead to substantially reduced mean squared errors compared with existing methods.

106 Efficient Data Collection Techniques for Cutting-Edge Applications ■●

Section on Physical and Engineering Sciences, International Indian Statistical Assoc., Reps. for Young Statisticians, Section on Quality and Productivity

Monday, August 1, 8:30 a.m.–10:20 a.m.

Adaptive Sequential Experimental Designs for Large-Scale Multiple Hypothesis Testing

✦Robert Nowak, University of Wisconsin-Madison, , *nowak@ece.wisc.edu*; Jarvis Haupt, University of Minnesota; Rui Castro, Eindhoven University of Technology

Key Words: multiple testing, experimental design

This talk discusses the role of adaptive experimental designs in the context large-scale multiple hypothesis testing problems, which are of central importance in the biological sciences today. Formally, consider p independent tests of the form H0: X ~ N(0,1) vs. H1: X ~ N(m,1), for m>0. It is well known reliable decisions are possible only if m, the signal amplitude, exceeds sqrt(2 log p), when p is very large. However, it can be shown that this limitation only exists because all the data are collected prior to testing. Distilled Sensing (DS) is an adaptive multistage experimental design and testing procedure that implements this refinement idea. Given the same experimental budget, DS is capable of reliably detecting far weaker signals than possible from non-adaptive measurements. It can be shown that reliable detection is possible so long as the signal amplitudes exceed any arbitrarily slowly growing function of p. For practical purposes, this means that DS is capable of reliable detection at signal-to-noise ratios that are roughly log(p) weaker than that required by non-adaptive methods.

Experiment Design and Optimization for Scenario Assessment Based on Computer Simulations

✦Brian Williams, Los Alamos National Laboratory, , *brianw@lanl.gov*; Christine Anderson-Cook, Los Alamos National Laboratory; Leslie M. Moore, Los Alamos National Laboratory; Jason Loeppky, University of British Columbia Okanagan; Cetin Unal, Los Alamos National Laboratory

Key Words: computer model, calibration, scenario assessment, experiment design, optimization, uncertainty quantification

We investigate experiment design and optimization strategies for scenario assessment with complex multi-physics codes. Component physics models (e.g. material strength, equation of state) contain uncertain parameters that are calibrated to experimental data. Other parameters allow for scenario definition, such as the geometry of the physical system or assumed operating conditions. We wish to identify regions of the scenario variable space in which at least one critical metric operates outside of its control bounds. Scenario assessment that properly accounts for uncertainty requires physics parameter uncertainty to be propagated through calculated performance metrics. It may be necessary to reduce this uncertainty through improved calibration of physics model parameters to expand the domain of acceptable scenario variation. Since we assume that multiple performance metrics will be utilized to determine the best new data to collect conditional on current understanding, we propose an approach that utilizes a comprehensive resource allocation framework that updates the current analysis and optimally reduces residual physics uncertainty and optimizes the regime of allowable scenarios.

Experimental Designs for Statistical Learning

✦Xinwei Deng, Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706 USA, *xdeng@stat.wisc.edu*; Peter Z. G. Qian, University of Wisconsin-Madison

Key Words: design of experiments, machine learning, cross-validation

This talk is devoted to showcasing the importance of design of experiments for machine learning. The basic message is if you can design better, then you can learn better. Specific topics include sliced cross-validation for estimating the error rate of a classifier, designs for the lasso and sliced designs for efficient tuning parameter selection. Joint work with Xinwei Deng at U of Wisconsin-Madison and C. Devon Lin at Queen's University, Canada.

107 Quantile Regression, Time Series and Extremes with Applications to Business Forecasting and Risk Management ■●

Business and Economic Statistics Section, International Indian Statistical Association, Section on Nonparametric Statistics, Section on Risk Analysis

Monday, August 1, 8:30 a.m.–10:20 a.m.

Recovering the Tail Shape Parameter of the Risk Neutral Density from Option Prices

✦Kamal Hamidieh, Rice University, Department of Statistics, Houston, TX 77251-1892 USA, *kh1@rice.edu*

Key Words: Risk Neutral Density, Options, Tail Shape Parameter, Generalized Pareto Distribution, Extreme Value Theory, Asset Pricing

In this paper closed form pricing formulas for the out of the money European style options are derived, and a method to recover the tail shape of the risk neutral density from the observed option prices is developed. The pricing formulas satisfy many well known model-free no-arbitrage bounds for the European style options. The method is quite general, and applies to a large class of risk neutral densities which also includes the lognormal density. The method is original and unique in the sense that the focus is only on the tails of the risk neutral density and not on the entire body of the density as many works have already done this. A case study using the S&P 500 option prices is given. In particular, the estimation of the tail shape of the S&P 500 index just prior to the market turmoil of the September 2008 shows a "thickening" of the left tail of the risk neutral density but "thinning" in midst of the turmoil. Information gained from this method would be useful to the risk managers, researchers, and practitioners interested in assessing and quantifying future extreme market conditions based on the observed option prices.

Sparse-Moving Maxima Models for Extreme Dependence in Multivariate Financial Time Series

✦Zhengjun Zhang, University of Wisconsin, 1300 University Ave, Statistics MSC 1220, Madison, WI 53706, *zjz@stat.wisc.edu*

Key Words: Extreme value theory, GMM estimator, Value-at-Risk, Max-stable processes

The multivariate maxima of moving maxima (M4) model has the potential to model both the cross-sectional and temporal tail-dependence for a rich class of multivariate time series. The main difficulty of applying M4 model to real data is due to the estimation of a large number of parameters in the model and the intractability of its joint likelihood. In this paper, we consider a sparse M4 random coefficient model (SM4R), which has a parsimonious number of parameters and it can adequately capture all the major stylized facts exhibited by financial time series found in recent empirical studies. We study the probabilistic properties of the newly proposed model and develop a new approach for statistical inference based on the generalized method of moment (GMM). We also demonstrate through real data analysis that the SM4R model can be effectively used to improve the estimates of the value at risk for portfolios consisting of multivariate financial returns while ignoring either temporal or cross-sectional extreme dependence could result in serious underestimate of market risk.

A Gini Autocovariance Function: Formulation, Properties, Estimation, and Applications

✦Robert Serfling, University of Texas at Dallas, , *serfling@utdallas.edu*

Key Words: Time series, Autocovariance, Gini covariance, Autoregressive models, Moving average models, ARMA models

A new type of autocovariance function, the "Gini autocovariance function", is formulated under merely first order moment assumptions. It plays roles similar to those of the usual autocovariance function which, however, requires second order assumptions. Key properties of this new tool and suitable estimators are discussed. Systems of equations for the parameters of autoregressive (AR), moving average (MA), and ARMA models entirely in terms of the Gini autocovariance function are derived. For the AR case, these are linear, yielding convenient, easily interpreted, closed form expressions. Simulation results comparing the "Gini" estimators with least squares estimators for these models are discussed. Without second order assumptions, least squares estimators still perform best but lack population analogues, whereas the Gini sample versions are competitive and do possess population analogues as long as first order assumptions hold. Additional perspectives will be provided.

Can Robust Quantile Regression Improve Valuation of Baseball Players?

✦Jonathan Lane, Rice University, Houston, TX , *jono.lane@rice.edu*; David W. Scott, Rice University

Key Words: Quantile Regression, Robust Estimation, Baseball, Player Valuation

In Major League Baseball, teams spend large amounts of time, effort, and money to determine the value of individual players. In recent years, statistical analysis has been used, in addition to traditional scouting methods, to perform this valuation. By using statistical analysis, teams can estimate the value of a player's past performance, as well as predict the value that the player will produce in subsequent years. Here, we examine the effects of using robust quantile regression, particularly using L2 estimation, to perform these prediction and compare them to currently used prediction methods.

108 Neyman Lecture ●

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

Monday, August 1, 8:30 a.m.–10:20 a.m.

Applied Bayesian Nonparametrics

✦Michael I. Jordan, University of California, Berkeley, , *jordan@cs.berkeley.edu*

Bayesian inference is often viewed as an assumption-laden approach to statistical inference, in which strong assumptions are imposed in order to support inference. The past two decades have seen the development of an increasingly vigorous field of Bayesian nonparametrics, which simultaneously provides an expressive language for prior specification and allows for weaker assumptions. Mathematically, Bayesian nonparametrics amounts to using general stochastic processes as prior distributions. I discuss a class of stochastic processes known as "completely random measures" that I view as providing a useful point of departure for a range of applications of Bayesian nonparametrics to problems in science and engineering. In particular I will present models based on the beta process, the Bernoulli process, the gamma process and the Dirichlet process, and on hierarchical and nesting constructions that use these basic stochastic processes as building blocks. I will discuss applications to a variety of scientific domains, including protein structural modeling, vision, speech and statistical genetics.

109 Statistical education: Learning what works ■

IMS, Section on Statistical Education

Monday, August 1, 8:30 a.m.–10:20 a.m.

Thinking Critically About Undergraduate Testing

✦Eric Loken, Pennsylvania State University, 119B Henderson Building, University Park, PA 16802, *loken@psu.edu*

Key Words: Teaching, Measurement, Testing

At many large universities, science, technology, engineering, and mathematics (STEM) departments function like major testing organizations, delivering hundreds of thousands of multiple choice tests. And yet departments rarely dedicate any resources to the construction, analysis or evaluation of tests. This is surprising considering risks in accountability, and lost opportunities for innovation in pedagogy. Multiple choice tests are often guaranteed to provide unequal information across the ability spectrum, and almost nothing is known about the consistency of measurement properties across subgroups. Course management systems that encourage testing from item banks can expose individual students to dramatically unequal assessment. Aside from issues of fairness and validity, the neglect of research on testing in undergraduate classes represents a missed opportunity to take an empirical approach to pedagogy. Years of testing have generated vast amounts of data on student performance. These data can be leveraged to inform pedagogical approaches. They can also be leveraged to provide novel assessments and tools to better encourage and measure student learning.

Modernizing the Introductory Statistics Course: Challenges and Rewards

Mark Hansen, University of California; ✦Deborah Nolan, University of California, Department of Statistics, 367 Evans Hall MC 3860, Berkeley, CA 94720-3860, *nolan@stat.berkeley.edu*

Key Words: education, introductory statistics, graphics, simulation

In our efforts to modernize the statistics curriculum, we have experimented with using R in lower-division courses for the non-major. Our goals are to expose students to the vast amount of digital information available to them and how statistics can be used to analyze and make sense of these data. We have found that incorporating R into our courses necessitates change in both what we teach and how we teach. We spend more time on graphics, where students create visualizations that tell stories and uncover structure in data. Also, students layer small computational steps to construct simulations to understand a statistical concept or assess whether data matches expectations (whether those expectations are for a plot or summary statistic). One challenge we have encountered is how this approach does not integrate well with traditional curriculum and textbooks that emphasize normal approximations over exact and simulated distributions, univariate over multivariate approaches, simple graphics such as stem-and-leaf plots over, e.g., density, mosaic, violin plots, etc. It is time to abandon the old curriculum and design a modern statistics course from scratch.

Statistics: What's the Difference? An Introductory Statistics Course

✦Vincent Dorie, Columbia University, NY 10022, *vincent@stat.columbia.edu*; Andrew Gelman, Department of Statistics, Columbia University; Valerie Chan, Columbia University

Key Words: introductory, education, R, active learning

In our experience, we have found wide-spread dissatisfaction with the content of traditional introductory statistics courses: students often build intuition separate from mathematical skills, complicating evaluation; unlike in physics or calculus, there is no natural progression to the topics; and many of our students come from non-technical fields, viewing the course as an interruption to their studies. From the educator's perspective, much time is spent preparing and executing the class which leaves little room to develop a cohesive framework. To address this, we have developed an introductory course that includes, freely available: minute-by-minute class schedules and slides; an electronic textbook; homework and example problems; an R package and sample code; and a test-bank of exam questions. Class time incorporates elements from the forefront of education research such as active learning, Peer Instruction (Mazur, 1997), and Just-in-Time-Teaching (Novak, 1999). In addition, we have reorganized and redefined the course content into a form that we believe increases coherence and relevance. Finally, we have implemented our program in classes containing over 100 students.

110 Beyond the Textbook ■

Section on Statistical Education, International Indian Statistical Association

Monday, August 1, 8:30 a.m.–10:20 a.m.

Beyond the Textbook

✦Webster West, Texax A&M University, Dept of Statistics, College Station, TX 77843, *west@stat.tamu.edu*; ✦Rebekah Isaak, University of Minnesota, 56 East River Road, Minneapolis, MN 55455, *isaak009@umn.edu*; ✦Christopher Barr, Harvard University, Dept of Biostatistics, Cambridge, MA , *cbarr@hsph.harvard.edu*

Key Words: textbooks, teaching, statistical education

In today's world a textbook that has been written and printed years before it reaches studdents taking a statistics course often containes unused material and rapidly become out of data. The internet and electronic media offer newer, customizable text materials for students, The panelists in this session will discuss new approaches to creating or using textbooks in teaching introductory statistics. Some of the alternatives to traditional textbooks disucssed incluce on-line texts, student-created texts, searchable Wikipedia-type texts, and collections of online readings.

111 Statistical Programmers in the Pharmaceutical Industry ■

Section for Statistical Programmers and Analysts, Biopharmaceutical Section

Monday, August 1, 8:30 a.m.–10:20 a.m.

Statistical Programming At Academic And Industrial Settings

✦Hsin-Yi (Cindy) Weng , Unviersity of Utah, Department of Pediatrics, 295 Chepeta Way RM 2S010, Salt Lake City, UT 84108 USA, *cindy.weng@hsc.utah.edu*

Key Words: Statistician, Academic Biostatistician, Biostatistican, statistical programming, FDA

Academic and Industrial Settings as a Statistician Hsin-Yi Cindy Weng1, Winning Volinn2, University of Utah Department of Pediatrics1, University of Utah Department of Family and Preventive Medicine2 There are three goals of Study Design and Biostatistics Center (SDBC) at University of Utah. SDBC promotes structural research by providing study design, analysis support and biostatistical education to researchers at University of Utah. We integrate existing biostatisitcal resources and biostatistical expertise through University Departments, which enhance the capacity of the center to specialize research areas, as well as provide access to general study support. At University research community, statisticians, epidemiologists and researchers are in a joint intellectual effort to enhance quality of research at the University. At academic setting, a biostatistician 's responsibility of statistical programming is not as rigorous as statistical programming at industrial environment comply with the FDA regulatory requirements and maintain an open communication channel with the FDA. However, knowledge and experiences in health sciences is essential.

The Future Of Statistical Analyst'S Work In The Pharmaceutical Industry: Trial Simulation

✦Natalie Cheung Hall, Eli Lilly and Co., 11797 Ledgestone Cir, Fishers, IN 46037, *cheungnw@lilly.com*

Key Words: Trial Simulations, Clinical Trial Design, Statistical Programming

In the pharmaceutical industry, the search for more efficient and safer trial designs requires increased use of trial simulations. These simulations allow project teams to vary components of a study in order to identify the benefits and risks of each scenario and arrive at the optimal trial design. However, the output from these simulations, whether performed through software or custom code, are voluminous and lack interpretability in their raw form. This talk will illustrate the statistical analyst's challenge of evaluating the data quickly and thoroughly, and communicating the operating characteristics of each trial scenario to the project team in order to make an informed decision about the study design. A number of key tools (e.g., simulation software, trial simulation plan template) that aid in this process will also be discussed.

Becoming A Pharmaceutical Statistical Programmer

Sandra Minjoe, Octagon Research Solutions; ✦Mario Widel, Roche Molecular Systems, Inc, , *mwidel@uicalumni.org*

Key Words: statistical programmer, inherent skill, acquired skill, skill gap

Is Pharmaceutical Statistical Programmer the right career for you? The answer can be at least partially determined by considering the set of inherent and acquired skills often used in this profession, and comparing these to your own skill base. The authors of this paper have been in the industry for a combination of 35 years and have observed many successful and not-so-successful pharmaceutical statistical programmers. This paper is a compilation of our experiences and expresses our opinions of the skills needed and best suited for the role of pharmaceutical statistical programmer. It can help you identify your skill gaps and describe what you can do about them.

Agile Approach To Overcome Evolving Challenges Faced By Statistical Programmers During Clinical Reporting

✦vikash jain, eClinical Solutions, 250 Michelle Lane, #209, Groton, CT 06340 u.s.a., *jainvikash77@yahoo.com*

Key Words: Statistical, Programmer, Agile, Clinical, Reporting, quality

Evolving industry standardization, global competitiveness and economic downturn are driving many pharmaceutical and biotech companies to look into strategic approaches to adopt business models of outsourcing or off shoring or blend of both for their clinical programming operations for reduced cost, maintain quality and minimal turnaround time on their reporting. With such strategic influences our paper will specifically focus on how the statistical programmer's involved in reporting phase are challenged when interacting with external cohesive departments and also challenged internally with in their programming team in operational aspects when most of them are dispersed domestically and/or globally. This paper outlines challenged scenarios faced at various avenues during their course of trial reporting either an individual study or a submission and how they can adopt to such situations proactively by incorporating continuous evaluation & adjustment, subjective judgment with agile thinking and understand their own shared purpose, goal and accountability with in their organization and also with external partnership to maintain the integrity of quality and relationship in a long run.

112 Progress in the use of Genomic Biomarkers in Clinical Trials ■

Biometrics Section, ENAR, International Chinese Statistical Association, Committee of Representatives to AAAS

Monday, August 1, 8:30 a.m.–10:20 a.m.

The Roles of the Cutoffs in Developing the Genomic Classifier for Preliminary Clinical Utility Assessment

✦Samir Lababidi, FDA, HFM-210, Rockville, MD 20852, *samir.lababidi@fda.hhs.gov*; Sue-Jane Wang, Office of Biostatistics, CDER, US FDA

Key Words: Gene Expression, Prediction performance, cutoff, MAQC-II, genomic classifier

With the advent of personalized medicine, microarray-based gene expression classifiers raise the hope of yielding diagnostic tests that help select optimum treatments for individual patients. It is critical to evaluate the gain in predictive ability using genomic classifiers over the clinical covariates. Depending on the cutoffs used for the genomic classifiers, the prediction performance of genomic models and clinical models can differ. In this talk, we assume the pre-specified clinical model is the model considered in practice. The performance of three candidate genomic models for the neuroblastoma study out of 25 models from the microarray quality control II (MAQC-II) project was examined. We compare the prediction performance of these genomic models with the clinical models, and further compare the genomic classifier combined with the clinical models to assess added value beyond the clinical covariates. Effective use of clinical covariates depends on the cutoff constructed for the genomic classifier. Any link of the genomic classifier with the biological pathway may be explored. These issues highlight the challenges in developing gene expression models.

Statistical Challenges in Genetic Biomarker Treatment Association with Clinical Outcomes: Experience from an Infectious Disease Phase II and Phase III Trials

✦Peggy H Wong, Merck, P.O. Box 2000, RY34-A316, Rahway, NJ 07901 US, *peggy_wong@merck.com*; Navdeep Boparai, Merck & Co., Inc.

Key Words: Pharmacogenetics, IL28B SNP, Phase II, feasibility

As the cost of genomic technology decreases, there has been a corresponding increase in the number of studies attempting to leverage genetics to assign treatments or dose. The results of these studies from literature, academic partnerships or internal studies can influence drug development real time. One challenge is to assess the feasibility and cost of qualifying that reported biomarker in ongoing phase II and III trials, or to define a biomarker in phase II for use in phase III. There are different challenges in using the same clinical trials for continued discovery. An infectious disease example will be used to drive the discussion. In this particular example, one IL28B SNP (rs12979860 on chromosome 19) was associated with 2-fold difference in sustained viral response in all patients with the CC genotype after administration of peg-IFN+RBV treatment in several studies (Fellay, 2010). The feasibility of incorporating that genetic research into the phase II and III studies based on assumptions on frequency of genetic variant, amount of consented population and expected effect. In addition, the results of the statistical testing of the genotypic association will be presen

Challenges in Developing and Validating Biomarkers That Are Treatment Induced: Experiences from an Oncology Phase III Trial

✦Bin Yao, Amgen, MS 24-2-A, One Amgen Center Drive, Thousand Oaks, CA 91320, *byao@amgen.com*; Yining Ye, Amgen

Key Words: oncology, biomarker, pharmacodynamic, validate, clinical utility

Unlike biomarkers identified at baseline, pharmacodynamic biomarkers are induced after treatment initiation. The utility of this type of biomarkers are often seen in early phase trials as a guide to dosing decisions and toxicity management. The apparent association between a pharmacodynamic biomarker and the clinical outcome may hold promises to identify patients who may or may not benefit from the treatment. However, there are major challenges to overcome in order to establish and to validate the clinical utility of such a biomarker. We will discuss several clinical trial design and analysis issues involving the development of a potential pharmacodynamic biomarker in oncology based on a phase 3 trial.

Identifying Predictive Biomarkers In Oncology Trials: A Practical Example From A Randomized Phase Ii Study In Neo-Adjuvant Breast Cancer (Bc)

✦Pralay Mukhopadhyay, BMS, 5 research pkwy, Wallingford, CT 06492, *pralay.mukhopadhyay@bms.com*; Guan Xing, Bristol Myers Squibb; Li-An Xu, BMS; David Liu, BMS; Christine Horak, BMS

Key Words: biomarkers, breast cancer, cutoff, enriched population

Identifying biomarkers that can differentially predict for response between experimental and comparator arms can be a key to success in oncology drug development. Such markers, if validated in Phase II settings could result in smaller, faster and more rationally deigned Phase III trials, conducted in selected patient populations. Here we discuss the design and analysis of a biomarker guided randomized phase II trial that was utilized to select potential biomarkers. The trial randomized 295 patients on a 1:1 basis to either paclitaxel or ixabepilone-based neoadjuvant therapy. Tissues samples were available for approximately 80% of the patients. One of the key objectives of the trial was to estimate an optimal cutoff for a pre-defined biomarker enriched population and estimate the response rates in that population. Logistic regression modeling was used to evaluate any interaction between treatment and biomarker. A five-fold cross validation method was used to estimate the optimal cutoff. The study failed to demonstrate presence of a marker that differentiates between treatment arms. Reasons why the trial failed to confirm the initial findings are discussed.

Robust And Reproducible Analysis Of A Very Large Clinical Genomics Study

✦John Storey, Princeton University, , *jstorey@genomics.princeton.edu*

Key Words: microarrays, longitudinal gene expression, R scripts

Trauma is the number one killer of individuals 1-44 years old in the United States. The prognosis and treatment of inflammatory complications in critically injured patients remains a great challenge, with a history of failed clinical trials and poorly understood biology. Over the course of several years, we have carried out an analysis of a large-scale study on 168 blunt-force trauma patients over 28 days, measuring ~400 clinical variables and longitudinal gene expression with ~800 microarrays. I will describe our analysis approach, which is complex and involves steps carried out by several researchers. I will then describe a "reproducible research" suite of R scripts that we have developed, which allows one to rapidly and transparently reproduce all results. The scripts may also easily be adapted to modify choices that we have made throughout the analysis pipelines. This is joint work with Keyur Desai, Chuen Seng Tan, and the Inflammation and the Host Response to Injury research program.

113 Health Policy Statistics Student Paper Competetion Winners

Section on Health Policy Statistics

Monday, August 1, 8:30 a.m.–10:20 a.m.

Health†Care†Reform And The Dynamics Of Uninsurance: Lessons†From†Massachusetts

✦John Graves, Harvard University, 12 Tufts Street, Unit 1, Cambridge, MA 02139 US, *jagraves@fas.harvard.edu*; Katherine Swartz, Harvard School of Public Health

Key Words: Health Reform, Health insurance, Health Policy

This paper provides the first dynamic look at changes in uninsurance spell durations in the wake of Massachusetts' 2006 health care reforms. Using longitudinal data from the Survey of Income and Program Participation, we utilize semi-parametric and non-parametric survival methods to estimate changes in uninsurance spell durations. We find evidence that insurance gains in Massachsuetts after 2007 were shared broadly among the short- and long-term uninsured. By contrast, we find no evidence that uninsured adults in two sets of northeastern comparison states had similar success in gaining coverage over the same time period. Non-parametric Kaplan-Meier estimates reveal that the distribution of uninsurance spells remained remarkably stable in comparison states between 2003 and 2009. In Massachusetts, we estimate a 3-month decline in the median uninsurance spell length and a decline of 11 months at the 75th percentile of spells. Our results suggest that under health insurance reforms that include an individual mandate, uninsurance becomes largely an acute event.

Accessing Diagnostic Accuracy With Ordinal Symptom Statuses Under The Absence Of A Gold Standard

✦Zheyu Wang, Department of Biostatistics, University of Washington, Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, *wangzy@u.washington.edu*; Xiao-Hua Andrew Zhou, Department of Biostatistics, University of Washington

Key Words: Diagnostic tests, EM algorithm, Volume under the ROC surface (VUS), Repeated ordinal data, Random effects models, Traditional Chinese Medicine (TCM)

Two common problems in assessing the accuracy of diagnostic tests or performance of doctors in detecting a symptom are the unknown true symptom status and the ordinal-scale of the symptom status. This is especially true for the studies on traditional Chinese medicine. In this paper, we proposed a nonparametric maximum likelihood method for estimating the accuracy of different doctors in detecting a symptom with an ordered multiple-class and without a gold standard. In addition, we extended the concept of the area under the ROC curve (AUC) to a hyper-dimensional overall accuracy measure and provided alternative graphs for displaying a visual result. The simulation studies showed that the proposed method has good performance in terms of bias and mean squared error. Finally, we applied our method to our motivating example on assessing the diagnostic abilities of five Chinese medicine doctors in detecting symptoms related to Chills disease. In addition, we discussed further on how to incorporate a dependence structure into the model under existence of a patient-level random effect. An ad-hoc test of the model fitting and a likelihood ratio test on the random effect were also provided.

Optimal Matching With Minimal Deviation From Fine Balance In A Study Of Obesity And Surgical Outcomes

✦Dan Yang, Department of Statistics, the Wharton School, University of Pennsylvania, 3730 Walnut St, Suite 400, Philadelphia, PA 19104, *danyang@wharton.upenn.edu*; Dylan Small, University of Pennsylvania; Paul R Rosenbaum, Department of Statistics, The Wharton School, University of Pennsylvania; Jeffrey H Silber, Center for Outcomes Research, The Children's Hospital of Philadelphia

Key Words: Assignment algorithm, Fine balance, Matching, Network optimization, Observational study, Optimal matching

In multivariate matching, fine balance constrains the marginal distributions of a nominal variable in treated and matched control groups to be identical without constraining who is matched to whom. In this way, a fine balance constraint can balance a nominal variable with many levels while focusing efforts on other more important variables when pairing individuals to minimize the total covariate distance within pairs. Fine balance is not always possible; that is, it is a constraint on an optimization problem, but the constraint is not always feasible. We propose a new problem that is always feasible and an algorithm which returns a minimum distance finely balanced match when one is feasible, and otherwise minimizes the total distance among all matched samples that minimize the deviation from fine balance. We also show how to incorporate an additional constraint. The case of knee surgery in the Obesity and Surgical Outcomes Study motivated the development of this algorithm and is used as an illustration.

Robust Inference In Semi-Parametric Discrete Hazard Models For Randomized Clinical Trials

✦Vinh Nguyen, University of California, Irvine, 8721 Lariat Ave, Garden Grove, CA 92844, *vqnguyen@uci.edu*

Key Words: Survival analysis, Censoring, Discrete time, Model misspecification, Robust inference

Time-to-event data in which failures are only assessed at discrete time points are common in many clinical trials. Examples include oncology studies where events are observed through periodic screenings such as CT scans. When the survival endpoint is acknowledged to be discrete, common semi-parametric methods for the analysis of observed failure times include the discrete hazard models (e.g., the discrete-time proportional hazards and the continuation ratio model) and the proportional odds model. In this manuscript, we consider estimation of an average treatment effect in discrete hazard models when the semi-parametric assumption is violated. Building on previous work for the continuous-time proportional hazards model we demonstrate that the estimator resulting from these discrete hazard models is consistent to a parameter that depends on the underlying censoring distribution. An estimator that removes the depedence on the censoring mechanism is proposed and its asymptotic distribution is derived. Basing inference using the proposed methodology allows for statistical inference that is reproducible and scientifically meaningful. Simulation is used to assess the proposed methodology.

Contrasting Evidence Within And Between Institutions That Supply Treatment In An Observational Study Of Alternative Forms Of Anesthesia

✦Jose R Zubizarreta, Department of Statistics, The Wharton School, University of Pennsylvania, , *josezubi@wharton.upenn.edu*; Mark Neuman, Department of Anesthesiology and Critical Care, The University of Pennsylvania School of M; Jeffrey H Silber, Center for Outcomes Research, The Children's Hospital of Philadelphia; Paul R Rosenbaum, Department of Statistics, The Wharton School, University of Pennsylvania

Key Words: Evidence factor, Fine balance, Optimal subset matching, Sensitivity analysis

A new matched design is proposed for contrasting evidence within and between institutions that supply treatment in an observational study. Using a new extension of optimal matching with fine balance, a first type of pairs exactly balance treatment assignment across institutions, so differences between institutions that affect everyone in the same way cannot bias this comparison. A second type of pairs compare institutions that assign most subjects to treatment and other institutions that assign most subjects to control, so each institution is represented in the treated group if it typically assigns subjects to treatment or alternatively in the control group if it typically assigns subjects to control. The design provides two evidence factors, that is, two tests of the null hypothesis of no treatment effect that are independent when the null hypothesis is true, where each factor is largely unaffected by certain unmeasured biases that could readily invalidate the other factor. The two factors permit separate and combined sensitivity analyses, where the magnitude of bias affecting the two factors may differ. The case of knee surgery with forms of anesthesia is considered in detail.

114 Statistical Modeling of Neural Spikes

■●

Section on Nonparametric Statistics

Monday, August 1, 8:30 a.m.–10:20 a.m.

The Pleasures And Pains Of Spike-Based Coding In The Nervous System: The Need For New Statistical Methodologies

✦Theodore W Berger, University of Southern California, Los Angeles, CA 90089 USA, *berger@bmsr.usc.edu*

Key Words: spike, brain, spatio-temporal pattern, nonlinear, dynamics

Neurons communicate with each other primarily with all-or-none bioelectrical events known as action potentials. Action potentials (spikes) provide for a pulse-based coding scheme, because the shapes of spikes are nearly equivalent, with only inter-spike-intervals (ISI) varying systematically in the presence of different events. Thus, neural representations in the brain are based on temporal patterns, or the sequence of ISIs. Because neural representations never rely on a single neuron, and instead emerge from the activity of a population of neurons, neural representations are based, not on temporal patterns alone, but on spatio-temporal patterns. Although "neural processing" is an often-maligned term, its meaning here is straightforward: it is the consequence of changing an incoming spatio-temporal pattern into a different, outgoing spatio-temporal pattern. Because the response of any neuron depends not only on the most recent input event, but also on the preceding history of activity, neural processing involves nonlinear transformations that can reach high orders. This session will discuss new statistical methodologies demanded by analysis of systems with such properties.

Modeling The Nonlinear Dynamics And Nonstationarities Underlying Spike Train Transformations In The Brain

✦Dong Song, University of Southern California, Department of Biomedical Engineering, Los Angeles, CA 90089-1451, *dsong@usc.edu*; Theodore W Berger, University of Southern California

Key Words: point processes, GLM, hippocampus, spike, prosthesis, Volterra model

Brain represents and processes information with spikes. To understand the biological basis of brain functions, it is essential to model the spike train transformations performed by brain regions. Such a model can also be used as a computational basis for developing cortical prostheses that can restore the lost cognitive function by bypassing the damaged brain regions. We formulate a three-stage strategy for such a modeling goal. First, we formulated a multiple-input, multiple-output physiologically plausible model for representing the nonlinear dynamics underlying spike train transformations. This model is equivalent to a cascade of a Volterra model and a generalized linear model. The model has been successfully applied to the hippocampal CA3-CA1 during learned behaviors. Secondly, we extend the model to nonstationary cases using a point-process adaptive filter technique. The resulting time-varying model captures how the MIMO nonlinear dynamics evolve with time when the animal is learning. Lastly, we seek to identify the learning rule that explains how the nonstationarity is formed as a consequence of the input-output flow that the brain region has experienced during learning.

Probabilistic Graphical Models Of Functional And Effective Neuronal Connectivity

Seif Eldawlatly, Electrical and Computer Eng. Dept., Michigan State University; Mehdi Aghagolzadeh, Electrical and Computer Eng. Dept., Michigan State University; ✦Karim Oweiss, Electrical and Computer Eng. Dept. and Neuroscience Program, Michigan State University, 2216 Engineering Building, East Lansing, MI 48824, *koweiss@msu.edu*

Key Words: graphical models, neuronal connectivity, spike trains

Coordination among cortical neurons plays an important role in mediating perception, cognition and motor actions. Deciphering the underlying neural circuitry is therefore of utmost importance in basic and clinical neuroscience. Probabilistic graphical models are powerful tools that can infer statistical relationships between the spiking patterns of simultaneously observed neurons. We present two methods for analyzing these patterns in multiple sensory systems based on graphical models. The first method employs dynamic Bayesian networks (DBNs) for inferring the effective connectivity among neurons. Unlike traditional pairwise correlation metrics, DBN accounts for precisely timed spiking of the entire neuronal population, allowing it to identify direct, possibly nonlinear, coupling between neurons by explaining away unlikely causes of firing. The second method identifies the functional connectivity between the neurons using a minimum entropy distance (MinED) method. MinED extends the well known maximum entropy models to model higher order interactions among neurons using hypergraphs. We demonstrate the use of both methods using data recorded from somatosensory and visual cortices.

Sparse Functional Model and Its Application to Neural Spike Activities

Yan Tu, Colorado State University; Jay Breidt, Colorado State University; ✦Haonan Wang, Colorado State University, Department of Statistics, Colorado State University, Fort Collins, CO 80523, *wanghn@stat.colostate.edu*

Key Words: Dynamic system, Group bridge, LASSO, Variable selection

In this talk, we consider the problem of modeling neural signal transformation. A dynamic Multiple-Input, Single-Output model of neural information communication is proposed. Each input neuron and the output neuron have a functional relationship which is approximated by polynomial splines. A penalized likelihood approach is implemented for simultaneous parameter estimation and variable selection. The notion of sparsity in parameter estimation has been generalized to function estimation. Two different types of functional sparsity are of particular interest: global sparsity and local sparsity. Computation of the penalized approach is rather challenging. The one-step estimator based on the group bridge approach for maximizing the penalized likelihood is proposed. The performance of the proposed method is assessed using Monte Carlo simulation studies.

Spike Train Kernel Methods For Neuroscience

✦Il Memming Park, University of Texas at Austin, Center for Perceptual Systems, 1 University Station, #A8000, Austin, TX 78712-0187, *memming@austin.utexas.edu*; Sohan Seth, University of Florida; Jose C. Principe, University of Florida

Key Words: point process, kernel method, spike train, neuroscience, characteristic kernel

Positive definite kernels has been widely used in the context of machine learning by the, so called, kernel machines such as the support vector machine and the kernel principal component analysis. An attractive property of a kernel machine is that it can be applied to arbitrary spaces as long as appropriate kernel is provided. We have developed spike train kernels and analyzed their properties in the context of two-sample problem, probability embedding as well as regression and classification. We discuss strictly positive definite kernels that provide theoretical foundation for its power.

115 Recent Advances in Statistical Matching ■●

Section on Survey Research Methods, Social Statistics Section

Monday, August 1, 8:30 a.m.–10:20 a.m.

Using Uncertainty Bounds For Regression Imputation In Statistical Matching

✦Hans Kiesl, Regensburg University of Applied Sciences, Fakultaet IM, Postfach 120327, Regensburg, International 93025 Germany, *hans.kiesl@hs-regensburg.de*; Susanne R‰ossler, Otto-Friedrich-Universit‰ot Bamberg

Key Words: statistical matching, regression imputation, uncertainty bounds

Statistical matching (also called data fusion) tries to combine information from different data sets by matching on those variables that are common to both files. Algorithms like nearest neighbour or Mahalanobis distance matching are routinely applied, but it is well known that they implicitly assume conditional independence of those variables that have not been jointly observed (called specific variables). In this paper, we discuss how to quantify the amount of uncertainty in the matching process by calculating bounds on distribution parameters of the specific variables. Data fusion might be viewed as a missing data problem, and we propose a regression imputation algorithm that creates different matched data sets with different feasible correlation matrices. Since several recent studies have used propensity score matching for combining different data sets, we will also discuss why propensity score matching is appropriate for the estimation of average treatment effects in the context of Rubin's causal model (where we have to deal with a different conditional independence assumption) but should not be applied in the data fusion setting.

Robust Multiple Imputation For Discrete Data In Missing-By-Design Settings

✦Florian Meinfelder, Universit‰ot Bamberg, Lehrstuhl f¸r Statistik und ÷konometrie, Feldkirchenstrasse 21, Bamberg, 96052 Germany, *florian.meinfelder@uni-bamberg.de*; Trivellore Raghunathan, University of Michigan

Key Words: multiple imputation, missing data, statistical matching, data fusion

Multiple Imputation (Rubin 1978, 1987) is a generally accepted method for analyzing incomplete data sets. Missing values are 'filled-in' or imputed m>1 times, thus creating m completed different data sets which are identical in the observed part, but vary over the imputed part. Most models and applications still focus on the imputation of continuous variables, and are usually based on normal distribution assumptions. However, survey data typically feature a large percentage of discrete data with non-normal distributions. This work addresses the multiple imputation of such variables in missing-by-design patterns (e.g. data fusion), where 'blocks' of data are missing. We propose Predictive Mean Matching (Rubin 1986) for vectors of variables as described by Little (1988) in combination with a Bayesian Bootstrap to create multiple imputations. An additional imputation step is needed, if the donor parts have missing values as well. This data situation can be seen as a mixture of missing-by-design with an overlaying ordinary item-nonresponse.

Bayesian Analysis Of Binary Probit Models: The Case Of Measurement Error And Sequential Regression Modeling For Missing Explaining Factors

✦Christian Aflmann, National Educational Panel Study, University of Bamberg, Wilhelmsplatz 3, Bamberg, 96047 Germany, *christian.assmann@uni-bamberg.de*

Key Words: MICE, MCMC, Bayesian Analysis, Probit, Measurement Error, Panel Data

The structure of multiple imputation algorithms is well suited for incorporation in MCMC estimation algorithms providing the analysis of primary interest. This paper implements two approaches to approximate the full conditional distribution of missing values within a sequential regression setup. In the context of a panel data set of bone ages with missing data, simple parametric models are chosen to provide an approximation of the full conditional distribution. Robustness checks are provided documenting the adequacy of the proposed approach. The resulting imputation algorithm is adapted within a MCMC algorithm allowing inference incorporating the uncertainty of missing explaining factors. Alternatively, a non parametric approach is chosen to mimic the full conditional distribution for missing values within variables with nominal and ordinal scale. This approach is applied within a Binary Probit Model incorportating a measurement error for the dependent variable aiming at an analysis of unit non response. Out-of-sample forecast criteria are used to gauge adequacy of non nested model spefications.

Statistical Matching Of Administrative And Survey Data - An Application To Wealth Inequality Analyses

✦Anika Rasner, German Institute for Economic Research (DIW Berlin), Mohrenstrafle 58, Berlin, International 10117 Germany, *arasner@diw.de*; Frick R. Joachim, German Institute for Economic Research (DIW Berlin); Markus M. Grabka, German Institute for Economic Research (DIW Berlin)

Key Words: statistical matching, imputation, inequality, social security wealth

Using population representative survey data from the German Socio-Economic Panel and administrative pension records from the Statutory Pension Insurance, the authors compare three imputation techniques (hotdeck, regression-based, univariate imputation sampling) and Mahalanobis distance matching to complement survey information on net worth with social security wealth information from the administrative records. The unique properties of the linked data allow for a direct control of the quality of matches under each technique. Based on three evaluation criteria (correlation coefficient, average distance, kernel density plots) comparing the observed and matched benefit, we identify Mahalanobis distance matching to perform best. Exploiting the advantages of the newly assembled data, we include social security wealth in a wealth inequality analysis. Despite its quantitative relevance in Germany, social security wealth has been thus far omitted because of the lack of adequate micro data. The inclusion of social security wealth doubles the level of net worth and decreases inequality by almost 25 percent. Moreover, the results reveal striking differences along occupational lines.

Displaying Uncertainty In Data Fusion By Multiple Imputation

✦Susanne R‰ssler, Otto-Friedrich-Universit‰t Bamberg, Feldkirchenstrafle 21, Bamberg, International 96045 Germany, *susanne.raessler@uni-bamberg.de*; Julia Cielebak, Otto-Friedrich-Universit‰t Bamberg

Key Words: statistical matching, bounds, missing data, combining data from different sources

Data fusion techniques typically aim to achieve a complete data file from different sources which do not contain the same units. Traditionally, data fusion, in the US also addressed by the term statistical matching, is done on the basis of variables common to all files. It is well known that those approaches establish conditional independence of the (specific) variables not jointly observed given the common variables, although they may be conditionally dependent in reality. However, if the common variables are (carefully) chosen in a way that already establishes conditional independence, then inference about the actually unobserved association is valid. Unfortunately, this assumption is not testable yet. Hence, we treat the data fusion situation as a problem of missing data by design and suggest imputation approaches to multiply impute the specific variables using informative prior information to account for violations of the conditional independence assumption. In a simulation study it is also shown that multiple imputation techniques allows to efficiently and easily use auxiliary information.

116 Non-Inferiority Trial Design Issues for Medical Devices ■

Biopharmaceutical Section

Monday, August 1, 8:30 a.m.–10:20 a.m.

Practical Issues In Non-Inferiority Trials

✦Chul H Ahn, FDA-CDRH, 10903 New Hampshire Avenue, Bldg 66, Rm 2218, Silver Spring, MD 20993, *chul.ahn@fda.hhs.gov*

Key Words: non-inferiority, superiority, non-inferiority margin, device trial, data quality, significance level

Non-inferiority trials are not simply underpowered superiority trials. In general, non-inferiority trials are more challenging to design, conduct and interpret. This talk will address some of the practical issues with non-inferiority trials such as selection of non-inferiority margin, sample size, significance level, testing for both superiority and non-inferiority, and data quality among others. Some specific examples of non-inferiority device trials will also be presented.

Non-Inferiority Tests For Binomial Endpoints In Two Arm Medical Device Trials

✦Maria Carola Alfaro, Boston Scientific, 4100 Hamline Avenue North, MS 9-315 South Town Square, Saint Paul, MN 55112-5798, *carola.alfaro@bsci.com*; Edmund McMullen, Boston Scientific

Key Words: Non-inferiority, binomial, sample size, type I error, power

Non-inferiority testing is performed in trials that intend to show that the new device is not inferior to the standard device. A non-inferiority test can be done on binomial and continuous endpoints. Our focus will be non-inferiority tests for the difference of binomial primary endpoints rates in two-sample trials. In this scenario a non-inferiority test is a one-sided test based on the binomial distribution. This test has impact on the sample size, type I error, power and non-inferiority margin of the trial in the design stage, and the conclusion of this test is based on the calculated one-sided p-value of the primary endpoint data. Through real case examples, we will analyze the impact of using different non-inferiority tests in the design stage and on the primary endpoint data through a comparison of estimations of sample size, type I error, power, non-inferiority margin and p-values.

Non-Inferiority Testing With A Variable Margin

✦Yonghong Gao, CDRH/FDA, 10903 New Hampshire Ave., Silver Spring, MD 20993, *yonghong.gao@fda.hhs.gov*

Key Words: Non-inferiority, variable margin, baseline covariates

The determination of non-inferiority margin is important in non-inferiority testing and the value of the margin is usually derived from historical data. If the margin depends on some baseline covariates, and the current active control population differs from the population used in historical trial in subject baseline covariates, namely violating the constancy assumption, then the non-inferiority margin is a variable margin. In this talk we will look at some analysis approaches to conduct non-inferiority testing under the covariate dependent margin.

Binomial Confidence Intervals For Testing Non-Inferiority Or Superiority: A Practitioner'S Dilemma

✦Vivek Pradhan, Boston Scientific Corp, 01752 USA, *vivek.pradhan@bsci.com*; John C Evans, Boston Scientific Corp; Tathagata Banerjee, IIMA

Key Words: coverage, expected length, location, non-inferiority, superiority

In testing for non-inferiority or superiority in a single arm study, the confidence interval of a single binomial proportion is frequently used. There are several intervals available in the literature, and most of them can be implemented using available software packages. In carrying out these tests, practitioners often face the serious problem that, different intervals may lead to conflicting conclusions, regarding acceptance or rejection of the null hypothesis. We address this question by investigating the performances of nine commonly used intervals of a single binomial proportion, in the light of three criteria, viz., coverage, expected length and location of the interval. An example with device trial data is presented.

Elicitation Of Non-Inferiority Margin

✦Xiting (Cindy) Yang, FDA-CDRH, 10903 New Hampshire Avenue, Silver Spring, MD 20993, *Xiting.Yang@fda.hhs.gov*

Key Words: elicitation, non-inferiority, device trial

In designing and analyzing non-inferiority trials, the choice of a non-inferiority margin always plays an important role in affecting trial characteristics. With little tool in use, such choice many times falls on the sample size determination and introduces many biases and difficulties in interpreting trial results. In this talk, using elicitation to systematically obtain a proper non-inferiority margin to help facilitate communication is discussed.

117 Spatio Temporal Models: Novel Applications ■●

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

Monday, August 1, 8:30 a.m.–10:20 a.m.

Wild Bootstrap For M Estimators Of Linear Regression

✦Xingdong Feng, National Institute of Statistical Sciences, 19 T.W. Alexander Dr, RTP, NC 27709, *xfeng@niss.org*; Xuming He, University of Illinois at Urbana-Champaign; Jianhua Hu, MD Anderson Cancer Center

Key Words: Wild bootstrap, M estimator

The wild bootstrap method is capable of accounting for heteroscedasticity in a regression model. However, the existing theory has focused on the wild bootstrap for linear estimators. In this note, we substantially broaden the validity of the wild bootstrap methods by providing a class of weight distributions that yield asymptotically valid wild bootstrap for M estimators of linear regression, including the least absolute deviation estimator and other regression quantile estimators. It is interesting to note that most weight distributions used in the existing wild bootstrap literature lead to biased variance estimates for nonlinear estimators of linear regression, and that for asymmetric loss functions a simple modification of the wild bootstrap admits a broader class of weight distributions. A simulation study on median regression is carried out to compare various bootstrap methods and to demonstrate the relevance of our work in finite-sample problems. With a simple finite-sample correction, the wild bootstrap is shown to be a valuable resampling method to account for general forms of heteroscedasticity in a regression model with fixed design points.

Bayesian Spatio-Temporal Syndromic Surveillance - An Epidemiological Prospective

✦Jian Zou, National Institute of Statistical Sciences, 19 T.W. Alexander Drive, P.O. Box 14006, RTP, NC 27709, *frankzou@gmail.com*

Key Words: Spatio-temporal, Syndromic surveillance, Markov random field, SIR model

Early and accurate detection of outbreaks is one of the most important objectives of syndromic surveillance systems. This talk proposes a general Bayesian framework for the syndromic surveillance system. The methodology incorporates Gaussian Markov random field (GMRF) and spatial-temporal CAR modeling. The disease transmission mechanism is modeled through a person-to-person contact basis. This is comparable to traditional epidemiological models such as the SIR model. The model has some nice probabilistic representations, as well as attractive statistical properties. We demonstrated that the model is capable of capturing early outbreaks through extensive simulation studies and synthetic data generated from a dynamic SIR model.

Estimates Of Uncertainty In Generalized Linear Regression Models With Spatially Misaligned Data

✦Kenneth Kyle Lopiano, University of Florida, 1710 NW 2nd Avenue, Apartment 14, Gainesville, FL 32603, *klopiano@ufl.edu*; Linda J Young, University of Florida; Carol A Gotway, Centers for Disease Control

Key Words: Spatial Misalignment, Berkson Error, Kriging, Generalized Linear Models

Researchers in the fields of climate change, environmental risk assessment, and public health often augment their data collection with existing data or work entirely with existing data from multiple sources. When the datasets have a spatial component, the datasets are often spatially misaligned. Spatial misalignment occurs when two or more variables are observed at different locations or aggregated over different geographical units. When the datasets are spatially misaligned, the data must be combined using a common set of geographical units before relationships can be assessed. When smoothing techniques, such as kriging, are used to align the disparate datasets, Berkson-type measurement error is induced. As a result, although regression estimates are unbiased, estimates of their uncertainty are biased. In this work, an iteratively reweighted generalized least squares approach is proposed that produces unbiased estimates of the regression parameter and its standard error when kriging is used to align datasets in point-to-point and point-to-areal misalignment problems. The statistical properties of the approach are presented and simulation studies are conducted.

Test Of Assumption Of Constant Relative Yield Total (Ryt) In Replacement Series Experiments

✦Nathan M Holt, Department of Statistics, IFAS, University of Florida, 415 McCarty C, PO Box 110339, Gainesville, FL 32611-0339, *nateholt@stat.ufl.edu*

In replacement series experiments with two species, X and Y, the Relative Crowding Coefficient (RCC) serves as an index of interspecies competition. Competition between X and Y is inferred when estimated RCC differs significantly from one. Tests of the null hypothesis that RCC=1 are predicated on the assumption that total yield of species X and Y in mixture is independent of relative planting frequency; that is, for a given number of plants, the total yield is constant whether only X, only Y, or a mixture of X and Y is planted. Estimated Relative Yield Total (RYT) significantly different from unity is evidence of planting frequency dependence. Methods to draw inference about RYT will be presented.

Habitat Prediction Of Large Mammals Using Satellite And Observed Presence-Absence Data

✦Michael Hyman, University of Florida, , *mhyman@stat.ufl.edu*

Key Words: Habitat Prediction, Predictive Modeling, Species Distributions

The distribution of animals over an area varies according to many factors, including habitat characteristics such as elevation, slope, distance to water, food availability, and species interactions. Species-distribution models have enhanced our understanding of the factors influencing species' density and habitat choices. These models use the association between environmental variables and presence-and-absence records for species over an area to predict habitat suitability and species distributions. In this analysis, a combination of remotely sensed datasets and field-collected data is used to predict habitat use by large mammals along the Chobe riverfront in Chobe National Park, Botswana. These data are used to make predictive habitat maps for the riverfront area. Logistic regression predicts habitat suitability of a specific area while a zero-inflated count model provides information about the species density given environmental factors. The resulting models can be used to direct management policies of the park and can be applied to conservation planning in other areas as well.

118 Statistical Inference for Mixture Models ●

IMS, International Indian Statistical Association, Section on Nonparametric Statistics

Monday, August 1, 8:30 a.m.–10:20 a.m.

Set Estimation In Locally Identifiable Mixture Models

✦Daeyoung Kim, Department of Mathematics and Statistics, University of Massachusetts Amherst, Lederle Graduate Research Tower 1434, Box 34515, Amherst, MA 01003-9305 United States, *daeyoung@math.umass.edu*; Bruce George Lindsay, Penn State University

Key Words: Confidence set, Finite mixture model, Identifiability, Likelihood

Statistical inference for the parameters in a parametric model can be done by the method of maximum likelihood which provides procedures for point estimation, set estimation and hypothesis testing. In the set estimation problem there are several literatures that suggest superiority of the likelihood confidence sets over the Wald confidence sets. It is practically important to develop guidance to assess the adequacy of using the Wald confidence sets for the parameters and Fisher information matrix for the ML estimator. In this talk we propose two diag-

nostics designed to assess difference between the likelihood set and the Wald set for the parameters in any parametric model, and show how one can adapt them to a finite mixture model where the topology of the mixture likelihood is complicated. These diagnostics are based on a set of samples simulated from the modal simulation (Kim and Lindsay, 2011) that efficiently reconstructs the boundaries of the targeted likelihood confidence sets.

A Mixture-Model Approach To Segmenting Magnetic Resonance Images

Wei-Chen Chen, Department of Statistics, Iowa State University; ✦Ranjan Maitra, Iowa State University, Department of Statistics, Iowa State University, Ames, IA 50011-1210 USA, *maitra@iastate.edu*

Key Words: magnitude Magnetic Resonance Images, initialization, Kolmogorov test, Rayleigh distribution

Magnetic resonance (MR) images are usually magnitudes of complex-valued intensities at a voxel. We develop a practical model-based approach to segmenting three-dimensional MR images using a mixture of Riceans rather than a mixture of Gaussians that has traditionally been assumed. Specifically, we develop a version of the expectation-maximization (EM) algorithm by introducing the discarded phase information at each voxel as missing observations. The complete likelihood is then a member of the regular exponential family, and the EM can then be implemented without the need for potentially unstable numerical optimization methods. Spatial context to the segmentation is also introduced in our mixture model. An added benefit of our approach above is the ready estimation of the variability in the estimate, which can potentially be used to provide quantification in our diagnosis. We evaluate our methodology on both realistic simulation datasets as well as images acquired on a physical phantom and on human subject.

A New Initialization Procedure For The Em Algorithm In Gaussian Mixture Models

✦Volodymyr Melnykov, North Dakota State University, Department of Statistics, North Dakota State University, Dept 2770, PO Box 6, Fargo, ND 58108-6050 USA, *volodymyr.melnykov@ndsu.edu*; Igor Melnykov, Colorado State University - Pueblo

Key Words: EM algorithm, Gaussian mixture model, initialization

The success of convergence of the EM algorithm in finite mixture models depends on effective initialization. There are multiple approaches proposed in literature that deal with this problem. However, there is no method that can be preferred over the others in all cases. We propose a new procedure for Gaussian mixtures that can be seen as a generalization of popular emEM and Rnd-EM algorithms. The suggested method demonstrates promising performance and good results in many cases. We illustrate the proposed approach on several simulated and classification datasets.

Beta Mixture with Application to Fitting the Empirical Distribution of P-Values

✦Bing Han, RAND Corporation, 1776 Main St, Santa Monica, CA 90034, *bhan@rand.org*

Key Words: Beta mixture, shape-restricted density, Bernstein polynomials, NPMLE

Beta mixture models are frequently used as a parametric approach to estimating densities with a bounded support. We further investigate the shape-restricted density estimation, e.g., both the density and its estimator are decreasing functions on (0,1). Toward this goal, we examine the classic beta-mixture models and Bernstein-type nonparametric approach. These models are applied to estimate the local false discovery rate in large scale simultaneous inference.

Mixture of Generalized Varying Effect Models

✦Xianming Tan, The Methodology Center, PSU, 204 E Calder Way, Suite 400, State College, PA 16801, *xzt1@psu.edu*; Runze Li, Penn State University

Key Words: mixture model, varying coefficient model, penalized spline, smoothing parameters

The mixture of varying coefficient models extend the classic varying-coefficient models to a mixture framework, by assuming an observation or a cluster of observations may come from one of several hidden sub-classes, and there is an unique varying coefficient model for each sub-class to describe the relationships between outcomes and covariates. We will discuss the estimation of such models using penalized splines. Key issues like the selection of smoothing parameters will be discussed. Both simulated data example and real data example will be presented to illustrate the performance of our estimation approach.

119 Parallel Computation for Bayesian Inference

Section on Bayesian Statistical Science, Section on Statistical Computing

Monday, August 1, 8:30 a.m.–10:20 a.m.

Direct Sampling

✦Paul Damien, University of Texas-Austin, 1 University Station - B6000, Austin, TX 78712-1178, *paul.damien@mccombs.utexas.edu*

Key Words: Rejection Sampling, Auxiliary Variable, Bernstein Polynomial

In recent years, Markov chain Monte Carlo (MCMC) methods have been used to provide a full Bayesian analysis when the posterior distribution of interest is both analytically intractable, and it is not known how to draw independent samples. In this paper, a non-MCMC approach to sampling from posterior distributions is developed and illustrated. Some sampling problems, now thought to be best handled by MCMC methods alone, are tackled efficiently via independent samples.

Generalized Direct Sampling for Large-Scale Hierarchical Bayesian Models

✦Michael Braun, Massachusetts Institute of Technology, 77 Massachusetts Ave., E62-535, Cambridge, MA 02139 USA, *braunm@mit.edu*; Paul Damien, University of Texas-Austin

Key Words: Parallel computing, Simulation, Direct Sampling, Hierarchical Models

Generalized Direct Sampling (GDS) is an estimation method for hierarchical Bayesian models that generates samples from a multidimensional posterior distribution. Like the Direct Sampling (DS) algorithm of Walker, et.al., but unlike MCMC, GDS samples are independent, so they can collected in parallel (taking advantage of computers with multiple processing cores), without concerns about autocorrelation and chain convergence. But unlike DS, GDS separates the estimation method from the model specification, so the efficiency of the sampler does not depend directly on the choice of the prior. Also, GDS resolves a numerical stability problem of DS that makes GDS more useful for large-scale hierarchical models with many thousands of parameters. Consequently, GDS has the potential to replace MCMC as the preferred estimation method for a broadly general class of Bayesian models, and can be a frequently used addition to the statistician's toolbox.

Issues In Bayesian Datamining

✦John Liechty, Penn State University, 409 BB, Univeristy Park, PA 16802, *jcl12@psu.edu*

Key Words: Markov chain Monte Carlo, Data Mining, Parallel Computing

One advantage of using the Markov chain Monte Carlo (MCMC) algorithm for inference of hierarchical Bayesian models on extreme size data sets is that the sampling algorithm and data can be distributed to multiple nodes of a high performance computer in such a way that draws from full conditional distributions can be accommodated by passing just summaries of relevant parameters, between the different nodes as opposed to passing the data between nodes. These modeling and computation approaches allow for complex inference on extreme data sets in an efficient manner - allowing for schemes that analyze the entire data set as opposed to random subsets. In addition to providing some illustrative examples and results of computer experiments, we discuss the theoretical, convergence properties of the parallelized MCMC algorithm when using an asynchronous scheme for sampling and sharing parameter values across multiple nodes.

Exploiting Scala'S Parallel Collections And Actors For Parallel Bayesian Computations

✦David B. Dahl, Department of Statistics, Texas A&M University, College Station, TX 77840, *dahl@stat.tamu.edu*

Key Words: Bayesian nonparametrics, Dirichlet process, multithreaded programming, parellel computing, random partition models, SMP

As microprocessor clock speeds stagnate and multicore processors proliferate, strategies for parallel computing are becoming increasing important for advances in Bayesian computation. This talk provides an overview of parallelize within a single Markov chain Monte Carlo run on a single computer using Scala's new parallel collections library and Scala's actors. Scala is a programming language with excellent support for both object-oriented and functional programming. Scala's support for concurrency allows the statistician to naturally exploit conditional independence with little programming or computational overhead. The talk shows how to adapt traditional MCMC code to exploit parallelize and demonstrates how to access this code from with R. Wall and CPU times are compared. The ideas demonstrated in the talk are illustrated in several Bayesian models.

120 2020 Census: How can the Cost and Complexity of the 2020 Census be Controlled? ■

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

Monday, August 1, 8:30 a.m.–10:20 a.m.

2020 Census: How Can The Cost And Complexity Of The 2020 Census Be Controlled?

✦Sally M. Obenski, U.S. Census Bureau, 4600 Silver Hill Rd., Washington, DC 20233 U.S.A., *sally.m.obenski@census.gov*; ✦Jennifer Hunter Childs, U.S. Census Bureau, 4600 Silver Hill Rd, Washington, DC 20233, *jennifer.hunter.childs@census.gov*; ✦James B. Treat, U.S. Census Bureau, 4600 Silver Hill Rd., Washington, DC 20233, *james.b.treat@census.gov*; ✦Frank A. Vitrano, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, *Frank.A.Vitrano@census.gov*

Key Words: census, cost, multi-mode, nonresponse, operations

This panel will explore reasons for the growing complexity and cost of conducting a census. It is intended to spark discussion about whether simplification is possible in a diverse country of over 300 million people in different types of geography. The discussion will address the following questions: What are potential operational designs for the 2020 Census and how could they reduce cost and complexity?; How does the number of response modes and operations increase the complexity of the census?; and Why is getting the last 5 percent enumerated so difficult, and is it necessary?. This panel will provide an overview of early planning of the 2020 Census, focusing on potential operational designs, including expanding self response modes, using administrative records for nonresponse follow-up, and moving to extreme automation and fewer operations. The cost versus benefit of minimizing versus maximizing the number of modes and number of operations will be discussed. The discussion will conclude with methods used to obtain enumerations from the last 5 percent of the population and the cost versus benefit of gaining the last few responses.

121 The National Children's Study: Expansion of Methods Research and Analysis

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

Monday, August 1, 8:30 a.m.–10:20 a.m.

The National Children'S Study: Expansion Of Methods, Research, And Analysis

✦Jessica Graber, Eunice Kennedy Shriver National Institute of Child Health and Human Development, 6100 Executive Blvd; Room 5C01, Bethesda, MD 20892-7510, *graberje@mail.nih.gov*; Margot Brown, Eunice Kennedy Shriver National Institute of Child Health and Human Development; ✦Angela DeBello, NORC at the University of Chicago, 55 East Monroe Street, 30th Floor, Chicago, IL 60603, *debello-angela@norc.org*; ✦Martin Barron, NORC at the University of Chicago, 55 East Monroe Street, 30th Floor, Chicago, IL 60603, *Barron-Martin@norc.org*; ✦Michael D, Sinclair, NORC at the University of Chicago, 4350 East-West Highway, 8th Floor, Bethesda, MD 20814, *sinclair-mike@norc.org*

Key Words: Children's Health, Longitudinal, Health Outcomes, Recruitment

The goal of the National Children's Study is to improve the health and well-being of children and contribute to understanding the role various factors have on health and disease. Longitudinal in design, the Study will gather data on a wide range of environmental exposures believed to affect child health and development. Recently, an Alternate Recruitment Substudy was launched to assess the feasibility, acceptability and cost of study logistics and three recruitment models. Thirty sites have joined the original 7 sites to comprise this pilot, called the NCS Vanguard Study. This panel will provide a study update, relate experiences and lessons learned thus far, discuss the current Alternate Recruitment Substudy methods, and discuss data collection and analytic activities underway. A description of a unique data repository system designed to aggregate frequent deliveries of questionnaire, metadata, paradata and cost data for analytic and field monitoring purposes will be described. Specific challenges and potential solutions for linking and integrating aggregate and individual data from environmental, administrative and survey based data sources to the NCS survey respondents will also be discussed.

122 Risk Prediction and Estimation

Biometrics Section, ENAR, Section on Risk Analysis, Section on Statistics in Epidemiology, Scientific and Public Affairs Advisory Committee

Monday, August 1, 8:30 a.m.–10:20 a.m.

Sparse Bayesian Semiparametric Predictive Modeling with Applications in Unobserved Dose-Response Prediction

✦Ben Haaland, Duke-NUS Graduate Medical School, Singapore, 8 College Road, Singapore, International 169857 Singapore, *benjamin.haaland@duke-nus.edu.sg*

Key Words: Bayesian predictive modeling, functional data analysis, low-dimensional representation, prior specification, sparsity, Gibbs sampling

Statistical scientists are often confronted with the task of making informed predictions which will potentially have substantial financial and ethical impact in situations which are simultaneously novel and expected to be similar to a few relatively well-studied scenarios. Many of these situations are distinguished from typical regression-based prediction by the importance of expert information and the combination of only a few similar scenarios with high dimensional inputs and outputs. Here, we focus on a particular version of this problem that presents itself in the pharmaceutical industry. Predicting both side effect and endpoint dose-responses before the initiation of a clinical trial has enormous ethical and financial importance in the pharmaceutical industry. A sparse Bayesian semi-parametric model for predicting unobserved clinical dose-response curves conditional on preclinical data, data from similar compounds, and prior knowledge is proposed. Posterior sampling is achieved through a computationally efficient Gibbs sampler, allowing straightforward incorporation into a risk assessment model. The model is applied to actual data from the pharmaceutical industry.

Modeling Of Susceptibility Genes For Cancer Risk Estimation In Family Studies

✦Chih-Chieh Wu, MD Anderson Cancer Center, Department of Epidemiology, Unit 1340, 1155 Pressler Street, Houston, TX 77030, *ccwu@mdanderson.org*; Louise C Strong, MD Anderson Cancer Center; Sanjay Shete, University of Texas M. D. Anderson Cancer Center

Key Words: statistical genetics, family studies, familial cancer, genetic epidemiology

Numerous family studies have been performed to assess the associations between cancer incidence and genetic and non-genetic risk factors and to quantitatively evaluate the cancer risk attributable to these factors. As more and more mutated genes and risk alleles have been discovered or identified over the past few decades, it becomes increasingly important to incorporate information on known susceptibility genotypes into cancer risk analyses. However, mathematical models that account for measured (known) susceptibility genes have not been explored in family studies. We have developed a method to precisely model measured susceptibility genes accounting for intra-familial correlation in hereditary mutation distribution and to simultaneously determine the combined effects of individual risk factors and their interactions. Our method accounts for measured hereditary susceptibility genotypes of the proband and each relative in a family. Our approach is structured for age-specific risk models based on Cox proportional hazards regression. We exemplified the method by analyzing various family data of Li-Fraumeni syndrome with TP53 germ-line mutations and genetic modifier of MDM2 SNP309.

Evaluation Of Biomarkers As Principal Surrogates For Time-To-Event Endpoints

✦Erin E Gabriel, University of Washington, 201 Galer St., # 370, Seattle, WA 98109, *eeg14@uw.edu*; Peter B. Gilbert, University of Washington and Fred Hutchinson Cancer Research Center

Key Words: Surrogate endpoints, Causal effects, Time-to-event, Vaccine efficacy trail

Since 2002 when Frangakis and Rubin proposed a new definition of a surrogate endpoint based on causal effects, a principal surrogate, several methods have been developed for the evaluation and comparison of biomarkers as principal surrogates. To our knowledge only one such method has been developed for a time-to-event outcome, based on a structural Cox proportional hazards model. We introduce surrogate evaluation methods for a time-to-event endpoint using Weibull structural risk models, which allow for the evaluation of waning surrogate

value in the presence or absence of treatment efficacy waning. The causal estimand of interest can be identified even with missing potential outcomes by standard causal inference assumptions plus additional assumptions described by Gilbert and Hudgens (08 Biometerics), together with augmented trial design features outlined in Follmann (06 Biometerics) of baseline predictors of the potential surrogate and placebo subject cross-over to active arm and measurement of the potential surrogate. The new methods are motivated by and illustrated with a simulation study of a proposed multi-vaccine arm placebo-controlled efficacy trial design.

Accelerometer-Based Prediction Of Activity For Epidemiological Research

✦Jiawei Bai, Johns Hopkins University, 615 N. Wolfe Street, Room E3038, Baltimore, MD 21205, *jbai@jhsph.edu*; Ciprian Crainiceanu, Johns Hopkins University; Brian Caffo, Johns Hopkins Department of Biostatistics; Thomas A Glass, Johns Hopkins University

Key Words: Signal Processing, Accelerometer, Acceleration, Activity Prediction

We introduce statistical methods to quantify and classify daily activity using accelerometers. The methods are complimentary to the current standard based on the Activities of Daily Living (ADL) questionnaire, which suffers fundamental problems, such as memory and selection bias. In particular, we introduce models to predict several types of activity based accelerometers attached to the subject's waist. We show that simple means and standard deviations of acceleration at each second can be effectively used to distinguish between active and sedentary periods. Analysis of spectrum of the acceleration is used to locate walking time periods. Validation is provided to quantify the model accuracy of prediction.

Flexible Regression Model Selection For Survival Probabilities Of An Absorbing Event

✦Rui Wang, State University of New York at Albany, 1 University Place, Rensselaer, NY 12144, *coevrww@omh.state.ny.us*; Gregory DiRienzo, State University of New York at Albany

Key Words: Survival Analysis, Cox model, Cross-validation, Missing data

We propose a strategy to flexibly model survival probabilities as a function of baseline covariates for an absorbing survival event such as death. This strategy requires modeling the censoring random variable, which is right-censored in this situation. We propose a flexible approach to modeling the censoring variable that re-estimates its distribution as a function of covariates at each event time among only those subjects at risk. This setting can be cast as a missing-data problem where augmented inverse probability weighted complete-case estimators can be calculated. Specifically, we obtain consistent and asymptotically normal estimators of regression coefficients and average prediction error for each working survival model at the given time points of interest, that are free from the nuisance censoring variable. The model selection strategy uses multiple hypotheses testing procedures that control a given error rate when comparing working models based on estimates of average prediction error. An extensive simulation study shows the methods generally perform well and they are illustrated with analysis of the popular Mayo Clinic trial of Primary Biliary Cirrhosis (PBC)

A Finite Mixture Model To Refine Risk Stratification In Multiple Myeloma

✦Pingping Qu, Cancer Research And Biostatistics (CRAB), 1730 Minor Ave Suite 1900, Seattle, WA 98101 USA, *pingpingq@crab.org*

Key Words: multiple myeloma, risk prediction, mixture model, survival

Many cancers are characterized by tremendous heterogeneity in clinical outcome following standard therapies such as in multiple myeloma (MM). Accurately characterizing such heterogeneity and identifying patients at different risk groups can help uncover disease subtypes and help physicians develop targeted therapies. In recent years, many risk stratification models have been developed using genomic data. At UAMS (the University of Arkansas for Medical Sciences), a genomic 70-gene risk model has successfully identified high risk groups of MM patients for relapse. Here we present a finite mixture survival model to refine the risk assignment of this 70-gene model in the hope of better classifying patients with risk scores near the boundary of high and low risk. We discuss results from applying the mixture model to a large, UAMS myeloma clinical trial data set.

Model Prediction Of The Length Of Cancer Prior To Diagnosis With Application To Cancer Registry Data

✦Diana L Nadler, University at Albany School of Public Health, One University Place, Rensselaer, NY 12144, *dn0320@albany.edu*; Igor Zurbenko, State University of New York at Albany

Key Words: Weibull distribution, conditional survival analysis, maximum likelihood estimation, cancer, delay in diagnosis

Survival analysis statistics in cancer research are often reported in terms of individual survival from the time of diagnosis; however, the true time malignant cancer cells developed in the body is unknown. It has been noted that some cancers do not present symptoms in early stages and are diagnosed at late stages of disease, significantly reducing the chance of survival. This indicates there is a window of time allowing for earlier detection and increased long-term cancer survival which is of great public health concern. The proposed method accounts for the conditional assumption the individual survived up to the time of diagnosis and estimates the time interval between disease onset and diagnosis for different cancer types. Using the fundamental memory property of the Weibull distribution and maximum likelihood estimation, this method provides a unique point estimate of the length of the hidden, unobserved period of cancer growth given a random sample of observed survival times. This method may slightly overestimate this period, but will undoubtedly allow us to order cancer types by increasing risk to identify the silent killers with long, undetected periods of growth.

123 Bayesian computation 1

Section on Bayesian Statistical Science

Monday, August 1, 8:30 a.m.–10:20 a.m.

Bayesian Optimal Sequential Design For Random Function Estimation

✦Marco A. R. Ferreira, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, *ferreiram@missouri.edu*

Key Words: Bayesian analysis, EMCMC, expected utility maximization, simulation-based algorithms

We develop a novel computational framework for Bayesian optimal sequential design for random function estimation. This computational framework is based on evolutionary Markov chain Monte Carlo, which combines ideas of genetic or evolutionary algorithms with the power of Markov chain Monte Carlo. Our framework is able to consider general models for the observations, such as exponential family distributions and scale mixtures of normals. In addition, our framework allows optimality criteria with general utility functions that may include competing objectives, such as for example minimization of costs, minimization of the distance between true and estimated functions, and minimization of the prediction error. Finally, we illustrate our novel methodology with applications to design of a computer model experiment and experimental design for nonparametric function estimation.

A Bayesian Approach To Inverse Problems Via Feynman-Kac Formula

✦Radu Herbei, The Ohio State University, 1958 Neil Ave, Columbus, OH 43210 United States of America, *herbei@stat.osu.edu*

Key Words: Bayesian inverse problem, Feynman-Kac, partial differential equations, diffusion process

In modern applied statistics, scientists often use physical models based on partial differential equations. Such PDEs typically do not have closed form solutions and one has to use a numerical scheme to approximate the solution over a regular grid. In the current work we make use of the Fenyman-Kac formula to provide an alternative to computationally intensive numerical schemes. We express the solution of a PDE in a probabilistic setting and then approximate it via Monte Carlo. We apply our method to a Bayesian approach to solving an oceanographic inverse problem.

Parallelized Langevin Hastings Sampling

✦Matthew M Tibbits, Pennsylvania State University, 333 Thomas Building, University Park, PA 16802, *mmt143@psu.edu*; Murali Haran, Pennsylvania State University; John Liechty, Penn State University; Benjamin Shaby, Statistical and Applied Mathematical Sciences Institute

Key Words: Langevin Hastings, Spatial GLMs, Dynamic Epidemic Modelling, Markov chain Monte Carlo, Parallel MCMC, Automated Tuning Procedure

Langevin Hastings (LH) algorithms are often utilized for multivariate distributions with strong dependence among the variables. In such cases, standard MCMC algorithms (e.g. random walk Metropolis-Hastings) typically result in slow mixing Markov chains. Using local information, the gradient and Hessian, the efficiency of the LH algorithm can be significantly improved. Because the gradient and Hessian often do not exist in closed form, we rely on numerical approximations, which can greatly slow down the modified LH algorithm. We demonstrate that parallel processing can mitigate the impact of the expensive approximations whilst preserving their benefit to the LH algorithm. Furthermore, while Markov chain Monte Carlo methods typically require the user to specify tuning parameters which vary for each model and every dataset considered, we construct an algorithm that automatically identifies reasonable values for LH based on a short initial run. We investigate the performance of the modified, parallelized LH algorithm within the context of two challenging examples: inference for a spatial generalized linear model, and a model for dynamic epidemic propagation across a social network.

Efficient Gaussian Process Models Using Random Projections

✦Anjishnu Banerjee, Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708, *ab229@stat.duke.edu*; David Dunson, Duke University ; Surya Tokdar, Department of Statistical Science, Duke University

Key Words: Gaussian Processes, Dimension Reduction, Subset Selection, Random Projections, Compressive Sensing

Gaussian processes are widely used in different domains, motivated in part by a rich literature on theoretical properties. A well known drawback of GPs that limits their use in many applications is the very expensive computation involved, requiring O(n^3) computations in performing the necessary matrix inversions with n denoting the number of data points. In large data sets, storage and processing also lead to computational bottlenecks and numerical stability of the estimates/predicted values degrade with n. To address these problems, various methods have been proposed, with recent options including predictive processes in spatial data analysis and subset of regressors in machine learning, but they are sensitive to the choice of knots and limited in estimating fine scale structure in regions that are not well covered by the set of knots. We propose an alternative random projection methodology. We demonstrate how our method is a generalization of previous approaches and connect hitherto unconnected approaches in Machine learning and statistics. We finally demonstrate superiority of the proposed approach from a theoretical standpoint and via simulated and real data examples.

Multiset Model Selection

✦Dipayan Maiti, Virginia Tech Dept of Statistics, 707 Appalachian Drive, Apt 2, Blacksburg, VA 24060 United States, *dipayanm@vt.edu*; Scotland Charles Leman, Virginia Tech

Key Words: Multiset, Bayesian Model Selection, Bayesian Model Averaging

The Multiset Sampler has previously been deployed and developed for e?cient sampling from complex stochastic processes. We extend the sampler and the surrounding theory to high dimensional model selection problems. In such problems e?cient exploration of the model space becomes a challenge since independent and ad-hoc proposals might not be able to jointly propose multiple parameter sets which correctly explain a new proposed model. In order to overcome this we propose a multiset on the model space to enable e?cient exploration of multiple model modes. The model selection framework is based on independent priors for the parameters and model indicators on variables. While under this method we do not obtain typical Bayesian model averaged estimates for the parameters, we show that the multiset model aver-

aged parameter estimate is a mixture a distribution from which the true Bayesian model probabilities and the model averaged parameter estimate can be obtained.

Reversible Jump Markov Chain Monte Carlo Analysis Of Multiple Changes In A Volcano'S Eruption Period

✦Jianyu Wang, Duke University, 222 Old Chemistry Building, Department of Statistical Science, Durham, NC 27708 United States, *jw163@stat.duke.edu*; Robert Wolpert, Duke University; James Berger, Duke University

We apply statistical modeling in the risk assessment of volcanic hazards. The goal is to investigate the changes in the eruption frequency of Montserrat's Soufriere Hills volcano and predict the probability of future catastrophic events. First, we will introduce the Reversible Jump Markov Chain Monte Carlo method. We will then describe a penalized mixture prior distribution developed to employ geologists' opinions and deal with consequent difficulty in the calculation of normalizing constants. In addition, we will present a simulation study that illustrates all these ideas. The overall results of the real data showed that our estimates coincide with significant geological changes of the volcano. Ultimately, the value of our approach lies in its ability to model point processes with multiple change points and its extension to other extreme natural hazards.

A Bayesian Approach For Joint Clustering Through A Dirichlet Process

✦Yubo Zou, University of South Carolina, 800 Sumter Street, Columbia, SC 29208, *zou@mailbox.sc.edu*; Hongmei Zhang, University of South Carolina; Wilfried Karmaus, University of South Carolina; Hasan Arshad, Southampton University and Hospital Trust

Key Words: Clustering, Bayesian method, Gibbs sampler, Metropolis-hastings method

The clustering problem has attracted much attention in the past few decades. Traditional approaches focus on the clustering of either subjects or variables. However, clusters formed through these approaches are possibly lack of homogeneity. We propose a clustering method through joint clustering. Specifically, the variables are first clustered based on the agreement of relationships between variable measures and covariates, and then subjects within each variable cluster are further clustered to form refined joint clusters. A Bayesian method is proposed for this purpose, in which a semi-parametric model is used to evaluate any unknown relationship between variables and covariates, and a Dirichlet process is utilized in the process of second-step subjects clustering. The major novelty of the method exists in its ability to produce homogeneous clusters composed of a certain number of subjects sharing common features on the relationship between some variables and covariates. We conduct simulation studies to examine the performance and efficiency of the proposed method, and apply it to methylation measures at multiple CpG sites and to Zernike coefficients measures of circular pupils.

124 Analysis of Highthrough-put Data

Biometrics Section

Monday, August 1, 8:30 a.m.–10:20 a.m.

The Application of Targeted Variable Importance Measurement in Dimension Reduction in Gene Expression Data

✦Hui Wang, Stanford University, 137 Crescent Ave, Sunnyvale, CA 94087, *hwangui@stanford.edu*; Mark van der Laan, University of California, Berkeley

Key Words: targeted maximum likelihood estimation, semiparametric model, variable importance measurement, dimension reduction

When a large number of candidate variables are present, a dimension reduction procedure is usually conducted to reduce the variable space before the subsequent analysis can be carried out. Inspired from the causal inference literature, we demonstrate that the variable importance measurement (VIM) based on targeted maximum likelihood estimation (TMLE) can be used for the purpose of dimension reduction. The TMLE-VIM is a two-stage procedure. The first stage resorts to a machine learning algorithm such as LARS and random forest. The second step improves the first stage estimation with respect to the variable of interest. Hence, TMLE-VIM enjoys the prediction power of machine learning algorithms, accounts for the correlation structures among variables, and at the same time produces more accurate variable rankings. When utilized in dimension reduction, TMLE-VIM can help to obtain the shortest possible list with the most truly associated variables.

The Doubly Contaminated Normal Model And Its Application To Microarray Data Analysis

✦Richard Charnigo, University of Kentucky, 851 Patterson Tower, University of Kentucky, Lexington, KY 40506-0027, *RJCharn2@aol.com*; Qian Fan, University of Kentucky; Hongying Dai, Children's Mercy Hospital

Key Words: mixture model, large scale hypothesis testing, contaminated normal model, contaminated beta model, multiple comparisons, gene expression

The contaminated beta model (Allison et al, 2002; Dai and Charnigo, 2008) or contaminated normal model (Dai and Charnigo, 2010) may be employed to describe the distribution of a large collection of P values or of the underlying Z test statistics, respectively, arising from a microarray experiment. The contaminated normal model has the advantage of explicitly accounting for the signs of the test statistics, which in a microarray experiment may correspond to gene overexpression (positive statistics) or underexpression (negative statistics). However, the contaminated beta model has the advantage of providing better fits to microarray data sets in which there are abundances of both overexpressed and underexpressed genes. Therefore, we propose a new doubly contaminated normal model to describe the distribution of a large collection of Z test statistics. The doubly contaminated normal model enjoys both of the advantages enumerated above. Point and interval estimators of parameters from the doubly contaminated normal model, along with related hypothesis testing procedures, are investigated theoretically and via simulations. We conclude with an application to a real microarray data set.

Combining And Comparing Multiple Serial Dilution Assays For Inferring Particle Concentration

✦Jarrett Jay Barber, Arizona State University, School of Mathematical and Statistical Sciences, PO Box 871804, Tempe, AZ 85287-1804, *Jarrett.Barber@asu.edu*

Key Words: card test, complement fixation (CF) test, complete spatial randomness, method comparison, homogeneous Poisson process, latent variables

In 1922, R. Fisher used a serial dilution assay to estimate the number of protozoa in a volume of soil---one of the first applications of maximum likelihood. In 1950, W. Cochran reintroduced essentially the same approach for estimating the "most probable number" of organisms in a liquid. More recent studies adopt a similar approach to estimating particle concentration. A serial dilution assay is an indirect method of measuring concentration, or count, of particles in solution whereby only the presence or absence of particles is measured for each dilution in a series of increasingly dilute solutions of the substance of interest. Direct measurement may be impossible or inconvenient. Common assumptions lead to the 'one-hit Poisson' model; it is assumed that single particles are detected---perfect detection. Estimation of particle 'intensity' follows straightforwardly. We discuss a model extension wherein we combine data from multiple types of serial dilution assays to compare the assays via their, now, estimated concentration detection thresholds. We illustrate with synthetic data and data on the bacterial disease, brucellosis, infecting elk in the Greater Yellowstone Ecosystem (GYE).

Identifiability Of Species Phylogenies Under The Coalescent Model

✦Laura Kubatko, The Ohio State University, 404 Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210, *kubatko.2@osu.edu*; Julia Chifman, The Ohio State University

Key Words: Phylogeny, species tree, coalescent theory, algebraic statistics, phylogenetic invariant

A phylogenetic tree is a graph that displays evolutionary relationships among a collection of organisms. The sequence data available for phylogenetic inference often include samples taken from multiple genes within each organism. This necessitates modeling of the evolutionary process at two distinct scales. First, given an overall phylogeny representing the actual evolutionary history of the species, individual genes evolve their own histories, called gene trees. Then, along each gene tree, sequence data evolve, leading to the observed data that is used for inference. The coalescent model provides the link between the evolution of the gene trees given the species tree, and the evolution of the sequence data given the gene trees. Phylogenetic invariants have been proposed as a tool for inferring phylogenies using data from a single gene, and their mathematical properties have been widely studied. In this talk, we consider the development of methods based on phylogenetic invariants developed specifically for species trees, as opposed to gene trees. In particular, we use methods from algebraic statistics to establish identifiability of the species phylogeny.

Some Theoretical Treatment In Spatial Prediction Of High-Frequency Monitoring Data

✦Xiaohui Chang, Department of Statistics, University of Chicago, Chicago, IL 60637, *xiaohui@uchicago.edu*; Michael Stein, University of Chicago

Key Words: space time modeling, wavelet analysis, fourier analysis, covariance function, time series analysis, non-stationarity

Fourier analysis has the difficulty of capturing the infrequent local changes that are often observed in meteorological data. We proposed a wavelet based approach to capture these sudden but strong local changes and illustrated covariance structure could become much simpler without sacrificing any accuracy nor uncertainty estimates in the predictions. In this paper, we'll present several theoretical problems arose in wavelet analysis and attempt to draw comparison between Fourier analysis and wavelet analysis in the context of space time modeling.

Uncertainty Propagation From Network Inference To Network Characterization

✦Weston Viles, Boston University, 111 Cummington Street, Boston, MA 02215, *wesviles@bu.edu*; Eric Kolaczyk, Boston University

Network-based data (e.g., from sensor, social, biological, and information networks) now play an important role across the sciences. Frequently the graphs used to represent networks are inferred from data. Surprisingly, however, in characterizing the higher-level properties of these networks (e.g., density, clustering, centrality), the uncertainty in their inferred topology typically is ignored. The distribution of estimators characterizing these networks defined implicitly through standard thresholding procedures can have distributions complicated by dependence inherent among the thresholded events. Motivated by this observation, we present a method by which the distribution of a sum of dependent binary random variables is approximated and demonstrate the method by exploring the problem of estimating network density - a simple but fundamental characterization of a network - in the context of correlation networks with Gaussian noise.

125 Federal Survey Estimate Quality: Some Issues and Some solutions ■

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

Monday, August 1, 8:30 a.m.–10:20 a.m.

Use Of Data From Different Data Collection Waves To Explore Potential For Non-Response Bias In The 2010 Namcs Mail Survey

Bill Cai, National Center for Health Statistics, CDC; ✦Qiyuan Pan, National Center for Health Statistics, CDC, 3311 Toledo Road, Hyattsville, 20782, *qap1@cdc.gov*

Key Words: Survey Methods, Mail, EMR, Item Response Rate, Non-response Bias

The National Ambulatory Medical Care Survey (NAMCS) is an annual survey of office-based physicians that gathers information about physicians and their practices through in-person interviews. A mail survey was conducted in 2010 on supplemental physician samples to enlarge the sample of physicians surveyed regarding the use of electronic medical records (EMR). Each physician in the mail survey was sent questionnaires at two week intervals and phone calls were attempted with non-respondents after the third mailing for a total of four data collection waves. Because both weighted and unweighted response rates for 2010 were only 68 and 66 percent, respectively, the potential for non-response bias will be investigated. For each wave of the survey, item response rates and responses for physician and practice characteristics such as EMR adoption status will be analyzed. This paper discusses the study methods and the results.

Exploring Nonresponse Bias In A National Health Expenditures Survey Of Institutions

✦Sameena Maryam Salvucci, Mathematica Policy Research, Inc., 600 Maryland Avenue, SW, Suite 550, Washington, DC, DC 20024-2512, *ssalvucci@mathematica-mpr.com*; Eric Grau, Mathematica Policy Research, Inc.

Key Words: Response Rates, survey estimates, nonresponse adjustment, Weighting, logistic regression, nonresponse modeling

We examined potential nonresponse bias in a new multi-mode national survey of mental health and substance abuse treatment facilities and its association with the response rate. We used data from the 2010 SAMHSA Survey of Revenues and Expenses (SSR&E), which was linked to data from the census of substance abuse facilities in the 2010 National Survey of Substance Abuse Treatment Services (N-SSATS) and the census of mental health facilities in the 2010 National Mental Health Services Survey (N-MHSS). We compared a range of facility characteristics of respondents and nonrespondents to SSR&E using chi-squared statistics. We also examined the nature and strength of the association between response rates and a set of independent characteristics using a multivariate logistic regression model.

Logistic Regression With Variables Subject To Post Randomization Method

✦Yong Ming Jeffrey Woo, The Pennsylvania State University, , *yjw102@psu.edu*; Aleksandra Slavkovic, The Pennsylvania State University

Key Words: Statistical disclosure control, logistic regression, generalized linear models, EM algorithm

An increase in quality and detail of publicly available databases increases the risk of disclosure of sensitive personal information contained in such databases. The goal of Statistical Disclosure Control (SDC) is to provide information in such a way that individual information is sufficiently protected against recognition, while providing society with as much information as possible, and needed for valid statistical inference. One such SDC method is the Post Randomization Method (PRAM), where values of categorical variables are perturbed via some known probability mechanism, and only the perturbed data are being released thus raising issues regarding disclosure risk and data utility. A number of EM algorithms are proposed to obtain unbiased estimates of the logistic regression model after accounting for the effect of PRAM. The effect of the level of perturbation and sample size on the estimates will be evaluated, and relevant standard error estimates will be proposed. The ideas will be extended to generalized linear models.

Bridging Livestock Survey Results To Published Estimates Through State-Space Models: A Time Series Approach

✦Stephen Busselberg, National Agricultural Statistics Service, 3251 Old Lee Hwy Room 305, Fairfax, VA 22030, *stephen_busselberg@nass.usda.gov*

Key Words: State-Space, Time Series, SARIMA, Survey Bias

The status quo in survey sampling is publication and dissemination of survey results which are directly calculated as a function of the sample design and the survey inclusion probabilities. This process is irreproachable for population parameter estimates for the case in which the only information known about the population of interest comes from the sample. In some cases, however, components of a system related to the population may be available from sources outside the survey sample. This external information may allow a deterministic solution to the sample survey items of interest at the population level. For this reason, the survey results may not be congruous with external data known to be extremely precise. The conflicting results may therefore be deemed inappropriate for publishing. Such is the case with livestock inventory surveys collected by the National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA). This paper details a solution to this issue of incongruity through the application of a State-Space model system.

Developing An Improved Jolts Item Imputation Approach

✦Mark Crankshaw, Bureau of Labor Statistics, 2 Massachussets Avenue, NE, Washington, DC 20212, *crankshaw_m@bls.gov*

Key Words: JOLTS, imputation

The Job Opening and Labor Turnover Survey (JOLTS) is a monthly Bureau of Labor Statistics survey that attempts to measure US labor market dynamism. JOLTS measures the number of job openings, hires and separations at the national, regional and industry level. Levels of these variables are of interest as well as implied employment change: the ratio of hires to separations. Both theory and empirical evidence from historical JOLTS data indicate that the ratio of hires to separations for JOLTS reporters is dependent upon whether the JOLTS reporter is expanding, stable, or contracting in terms of employment. While the current JOLTS imputation nearest neighbor algorithm may impute appropriate hires and separations levels, the current imputation algorithm is ill-suited to impute appropriate ratios of hires to separations. The level of reported employment reveals nothing about the employment dynamics of the establishment. A linear regression imputation algorithm that adequately accounts for the appropriate employment dynamic of the imputed unit will be proposed. Using historical JOLTS data, the efficacy of the new imputation will be contrasted with the current algorithm.

126 Novel analytical solutions to health outcomes

Section on Health Policy Statistics, ENAR, Section on Health Policy Statistics, Scientific and Public Affairs Advisory Committee

Monday, August 1, 8:30 a.m.–10:20 a.m.

Effect Of Health Workforce Distribution On Cesarean Rate

✦Imam Xierali, AAFP, 1350 Connecticut Ave, Suite 201, Washington, DC 20036, *ixierali@aafp.org*; Catherine Livingston, Oregon Health & Science University

Key Words: Cesarean rate, Certified Nurse Midwife, Obstetrician/Gynecologist, health workforce, spatial regression, United States

Cesarean rate has been steadily increasing in the U.S. and it varies regionally. This study examined the relationship between cesarean rate and health workforce distribution in the U.S. The 3-year (2003, 2004, and 2005) average cesarean rate by residency county of the mothers was regressed on health workforce variations and other potential county socioeconomic confounders. Cesarean rate and maternal health care workforce showed significant spatial dependence with the highest rates observed in southeast and southern U.S. Cesarean rates were significantly and negatively associated with Certified Nurse Midwife availability ($p < .0001$). However, Obstetrician/Gynecologist availability showed a positive association with cesarean rate ($p = .0144$). Moreover, the availability of Family Physicians and General Internists did not show significant association with cesarean rates. The significant spatial coefficient suggests that cesarean rate in a county was not only related to cesarean rates in nearby counties, but also related to maternal workforce availability within the county and across adjacent counties. A regional approach would be proper to address rising cesarean rate in the nation.

An Evaluation Of Clinical Outcomes Of Liver Transplant Recipients And Impact Of Clinical Indicators

✦Alexia Melissa Makris, University of South Florida, Department of Epidemiology and Biostatistics, 13201 Bruce B Downs Blvd., MDC-56, Tampa, FL 33617-3805 USA, *amakris@health.usf.edu*; Yiliang Zhu, University of South Florida

Key Words: Cox Proportional Hazard Model, race and ethnicity, failure time data, liver transplantation, time-varying covariates, disparities

Liver transplantation is accepted as the therapeutic option of choice for the individuals with end-stage liver disease. Although it provides reasonable survival, the outcomes of transplantation can vary and are dependent on several patient and non-patient related factors. We focus on liver transplant candidates registered nation-wide between 2002 and 2009. Potential determinants include clinical prognostic and demographic factors. Cox proportional hazards models were used. Potential determinants include clinical prognostic and demographic factors. Cox proportional hazards models were used to quantify the waiting time and survival time, and also to screen the determinants which included a large number of demographic, clinical, and prognostic factors. While we examine the disparity factor of race, we adjust for known covariates such as age, gender, diagnosis, diabetes, body mass index (BMI), hospitalization status at listing, receipt of dialysis, albumin and prior malignancy which was already established in the literature as significant. Under the organ allocation scheme dominated by the MELD score, variation in waiting time remained across regions and race.

How Does A Stochastic Graph Model Help To Control An Infectious Disease Outbreak?

✦Yasaman Hosseinkashi, University of Waterloo, 200 Candlewood Cr., Waterloo, ON N2L5Y9 Canada, *yhossein@math.uwaterloo.ca*; Shojaeddin Chenouri, University of Waterloo; Yasaman Hosseinkashi, University of Waterloo

Key Words: Stochastic graph process, contact network, measles outbreak, stochastic epidemic models, basic reproductive number

A stochastic graph model is a bridge between standard SIR or SEIR epidemic models and graph theory. This combination provides a framework for analyzing the propagation of an infectious disease over a susceptible population with known or estimated contact network. The model permits the estimation of a sequence of outdegrees and indegrees which quantify the role of individuals in the infection spread. In this work we use the out- and indegree sequence to detect the super spreaders and the resistant individuals in the data from 1861 measles outbreak in Hagelloch, Germany. The result is applied to the adoption of optimum movement bands or quarantine. The disease outbreak is simulated in the same population, under different control strategies. In addition to the detection of influential individuals, the fitted model also permits estimating a dynamic version of the basic reproductive number over the outbreak period. A time series plot of this index is applied in comparing the simulation results.

Repeated Lifetime Traumatic Events

✦Haekyung Jeon-Slaughter, The University of Oklahoma Health Sciences Center, 920 Stanton L. Young Blvd, WP 3212, Oklahoma City, OK 73104, *hattie-jeon-slaughter@ouhsc.edu*; Carol S North, VA North Texas Health Care System and The University of Texas Southwestern Medical Center; Phebe M Tucker, The University of Oklahoma Health Sciences Center; Betty Pfefferbaum, The University of Oklahoma Health Sciecences Center

Key Words: Repeated events, Traumatic events, Stochastic process

Anyone could experience multiple traumatic events during the life course. However, these seemingly independent events occur to certain people repeatedly and sometimes too often. Our study investigates whether these seeming uncorrelated traumatic events are actually correlated or dependent each other, whether environmental factors explain frequencies of traumatic experiences, and whether first time traumatic event experience increases risk of multiple traumatic event experience later. The lifetime traumatic events are death or threatened death, actual or threatened serious injury, and actual or threatened sexual violation.This study applies existing stochastic process models to test our study questions using three data sets and they are National Comorbidity Survey-Replication collected data from U.S. general population, 1995 Oklahoma City Bombing Survivors data collected from the specific population who experienced a truly random traumatic event, and Hurricane Katrina evacuee data collected from the population at risk

who experienced a predicted traumatic event. The findings support low income and minority population being at high risk of multiple lifetime traumatic experiences.

Restricted Mean Models For Transplant Benefit And Urgency

✦Fang Xiang, University of Michigan, Department of Biostatistics, 1415 Washington Heights, Ann Arbor, MI 48109, *xiangf@umich.edu*; Susan Murray, University of Michigan, Department of Biostatistics

Key Words: Dependent censoring, Pseudo observation, Restricted mean life, Survival, Time-dependent covariates, Transplant benefit

National lung allocation policy relies on statistical estimation of each individual's urgency and transplant benefit in defining a lung allocation score (LAS),both of which require accurate estimation of waitlist days lived.Risk factors are available to estimate patient urgency at their listing time,with more urgent patients removed from the waitlist as they either die or get transplanted.LAS is highly linked to both a patient's survival time and censoring (transplant) time. Therefore,it is crucial to adjust for dependent censoring in modeling estimated days of life. For estimation relevant to modeling an individual patient's urgency, we develop a model for the restricted mean as a function of covariates, using pseudo observations that account for dependent censoring. Simulation results show that our method performs well in situations comparable to the lung waitlist setting. A restricted mean model is also used to estimate days lived post-transplant based on individual risk factors at the time of transplant. The difference in LAS score for an individual, when properly accounting for dependent censoring, has high impact on the priority and timing of an organ offer for these patients

Hypertension In Young Adults: Is It Associated With Social And Behavioral Aspects Of Their Wellbeing

✦Soma Roy, Cal Poly, 1 Grand Avenue, Statistics Department, San Luis Obispo, CA 93407, *soroy@calpoly.edu*; Karen McGaughey, Cal Poly; Ann Yelmokas McDermott, Cal Poly

Key Words: texting, exercise, BMI, college students, video games, TV viewing

In this paper, we investigate the association between the incidence of pre-hypertension and hypertension in young adults, and social and behavioral aspects of their wellbeing, such as, whether they typically eat alone, how much they text, how much TV they watch, whether and how much they exercise, play video games, and gamble. We use data collected on a pilot study of 432 students at California Polytechnic State University. This dataset is especially interesting, providing a glimpse into the health and behaviors of college students. The dataset contains information on the usual demographics such as, age, sex, race/ethnicity, as well as information on body mass index, and blood pressure (BP), cigarette and marijuana use, and other variables of interest. Initial exploration shows a significant difference in incidences of pre-hypertension and hypertension between males and females ($p < 0.001$), with only 35% of males having normal BP, compared to 68% of females. The objective of this research is to not only understand the association in question, but to also increase awareness on campuses such as ours, so that we may be able to help our students be healthy as well as wise.

Built Environment And Obesity Risk Factors: Do Where You Live, Work And Commute Influence Your Weight Status? A Spatial Analysis Of Elementary School Personnel In New Orleans, La

✦Adriana Cristina Dornelles, Tulane University, 1440 Canal St, room 2021, New Orleans, LA 70112 United States, *adornell@tulane.edu*; Rice Janet, Tulane University; Webber Larry, Tulane University; Diego Rose, Tulane University

Key Words: multilevel, BMI, obesity, neighborhood, food business, food environment

Obesity has become a national concern and has reached epidemic proportions. Environmental factors may contribute to the increasing prevalence of obesity. To date, most of the studies assessing the relationship between weight and aspects of food environment have focused on one environment at time: home or work site. However, on a daily basis, most individuals experience more than one environment. People are exposed to their neighborhood and worksite environments and to the surrounding areas of their back-and-forth trajectory to work. Due to the fact that each of these environments may have several food outlets, it seems plausible that all of those environments will have an effect on the individual's body mass index (BMI). This paper explored the impact of those three environments, separately and together, on people's BMI. We used data from elementary school employees in the New Orleans metropolitan area and Dunn & Bradstreet data for the food businesses. A cross-sectional design was used to associate the built environments and BMI. Spatial and multilevel analysis was utilized in order to explore the impact of predictors at the individual and environmental levels.

127 Functional Data Analysis ■●

Section on Nonparametric Statistics, Section for Statistical Programmers and Analysts, Section on Quality and Productivity

Monday, August 1, 8:30 a.m.–10:20 a.m.

Functional Sliced Average Variance Estimates

✦Wenting Fan, University of Georgia, The University of Georgia Department of Statistics, 101 Cedar Street, Athens, GA 30602, *wtfan@uga.edu*; Yehua Li, University of Georgia

Key Words: Functional data analysis, Sliced average variance estimates, B-Spline

We address the problem of extending multivariate dimension reduction methods on functional data analysis. In this paper, Sliced Average Variance Estimates (SAVE) is considered for functional data where the response variable is a scalar and prediction variables are curves. Due to this nature, the estimation procedure of SAVE cannot be directly applied although some properties can be retained. To address this problem, we propose B-Spline estimator for conditional variance on functional data. We demonstrate that on a benchmark classification task for functional data, the classification result based on Support Vector Machine (SVM) with our extended SAVE and a non-linear form classifier achieves the best result, whereas a linear form classifier fails, while also outperforming the methods without dimension reduction.

Robust Multivariate Functional Principal Component Analysis

Pallavi Rajesh Sawant, Auburn University; ✦Nedret Billor, Auburn University, Department of Mathematics and Statistics, 364C Parker Hall, Auburn, AL 36849, *billone@auburn.edu*

Key Words: Functional data, Functional outliers, Functional principal components, Multivariate stochastic process, Robustness

Technological advances have facilitated the acquisition of multivariate stochastic processes, also known as random fields, a real-valued process, which is a function of a multidimensional variable. Although multivariate stochastic processes have received less attention than univariate processes they play an important role in many areas like fRMI studies and spatial statistics. Univariate functional principal component analysis is a key technique to represent the curves in lower dimension. In real life noise is ubiquitous in the random fields and they tend to bias the interpretation and conclusion of any study in an uncontrolled manner. The main purpose of this paper is to propose robust multivariate functional principal component analysis to identify atypical observations and reduce the dimensionality of multivariate functional data. Simulated data and a real data example are used to illustrate the effectiveness of the procedure.

Identifiability And Estimation Of Time-Varying Parameters In Ode Models With Applications In Viral Dynamics

✦Hongyu Miao, University of Rochester, 601 Elmwood Ave, Box 630, Rochester, NY 14642, *hongyu_miao@urmc.rochester.edu*; Hua Liang, University of Rochester; Hulin Wu, University of Rochester

Key Words: Ordinary Differential Equation, Time-varying Parameter, Identifiability, Spline-enhanced Nonlinear Least Square, Viral Dynamics

Ordinary differential equations (ODE) are prevailing modeling tools in investigations of viral dynamics such as HIV and influenza virus. Due to the complexity in the inverse problem for nonlinear ODE models, investigators usually consider constant parameters in their models. However, key kinetic parameters in ODE models could have a time-varying nature such that constant coefficients cannot accurately represent the dynamic interactions between model components. In this study, we investigated the identifiability and estimation techniques for time-varying parameters in ODE models. In particular, we combined the implicit function theorem and the differential algebra method to address the identifiability issue and we used the spline-enhanced nonlinear least squares (SNLS) approach to accurately estimate both constant and time-varying parameters in HIV and influenza infection models. We then verified the validity of our approaches via simulation studies. Our methods were also applied to clinical and laboratory data and resulted in interesting biological findings.

Semiparametric Functional Linear Model

✦Dehan Kong, North Carolina State University, 3002G Kings Ct, Raleigh, NC 27606, *dkong2@ncsu.edu*; Fang Yao, University of Toronto; Hao Helen Zhang, North Carolina State University

Key Words: Functional data analysis, Functional linear model, Model selection, Principal components, SCAD, Semiparametric

We propose and study a new class of semiparametric functional regression models. With a scalar response, multiple covariates are collected, a large number of which are time-independent and a few may be functional with underlying processes. The goal is to jointly model the functional and non-functional predictors, identifying important scalar covariates while taking into account the functional covariate. In particular we exploit a unified linear structure to incorporate the functional predictor as in classical functional linear models that is of nonparametric feature. Simultaneously we include a potentially large number of scalar predictors as the parametric part that may be reduced to a sparse representation. We propose an iterative procedure to perform variable selection and estimation, by naturally combining the functional principal component analysis (FPCA) and the smoothly clipped absolute deviation (SCAD) penalized regression under one framework. Theoretical and empirical investigation reveals that the efficient estimation regarding important scalar predictors can be obtained and enjoys the oracle property, despite contamination of the noise-prone functional covariate.

Functional Component Selection In Functional Additive Models

✦Hongxiao Zhu, SAMSI, , *hzhu@samsi.info*; Fang Yao, University of Toronto; Hao Helen Zhang, North Carolina State University

Key Words: Component Selection, Model Selection, Additive Models, Functional Data Analysis, Smoothing Spline

Functional additive model (FAM) provides a flexible framework to model the relationship between the responses and functional predictors, and its additive structure naturally overcomes the curse of dimensionality. The problem of functional component selection in FAMs is an important issue but less studied in literature. In this work, we propose a new regularization framework for joint model estimation and functional component selection in the context of Reproducing Kernel Hilbert Space (RKHS). The proposed approach takes advantage of the uncorrelated structure of the functional PCA scores, which greatly facilitates the implementation of the approach. Asymptotic properties of the new estimator are studied, and extensive simulation studies are performed to assess the performance of the new method. We finally apply the method to a real data set.

Analysis Of Longitudinal Data With Multiple Response Functions

✦Jeng-Min Chiou, Academia Sinica, 128 Sec 2 Academia Road, Nankang, Taipei, International 11529 Taiwan, *jmchiou@stat.sinica.edu.tw*

Key Words: Functional data analysis, Linear manifold, Traffic flow analysis, Varying coefficient functions

We propose a linear manifold modeling method of exploring dependency relationship between multiple random processes in longitudinal data. The linear manifold is defined through a set of data-determined linear combinations for the multiple component trajectories, subject to the condition that their variances are relatively small. The model is characterized by a set of varying coefficient functions under orthonormality constraints, leading to time-varying relationships between the multiple functional components. Under mild conditions, the integral of the linear manifold model variances can be expressed in a quadratic form, which facilitates the construction of the model. This linear man-

ifold modeling approach provides a tool for determining a set of linear relationships that govern the components of multiple random functions, and further yields noise-reduced multivariate component trajectories. The proposed approach is illustrated through an application to highway traffic flow analysis, where the linear manifold describes the relationships between the multiple functional measurements.

128 Topics in Experimental Design ■●

Section on Physical and Engineering Sciences, Section on Quality and Productivity

Monday, August 1, 8:30 a.m.–10:20 a.m.

Comparison Of Some Series Of Binary And Non-Binary Balanced Nested Row-Column Designs For Correlated Errors

✦Nizam Uddin, University of Central Florida, Department of Statistics, Orlando, FL 32816, *nuddin@mail.ucf.edu*

Key Words: Binary block, Nested row-column design, Correlated errors, Non-binary block

This paper looks at block designs with nested rows and columns in both binary and non-binary blocks that are balanced for neighbors in rows, in columns, and in diagonals when blocks are assumed to be on cylinders. Planar versions of these neighbor balanced cylindrical block designs are compared using some statistical optimality criteria under some error covariance structures. Similar to uncorrelated errors, non-binary designs are found to be more efficient than the corresponding binary designs for correlated errors.

Row-Column Designs with Minimal Units

✦Xianggui Qu, Oakland University, 2200 N. Squirrel Road, Rochester, MI 48309, *qu@oakland.edu*

Key Words: Binary design, (M,S)-optimal design, Row-column design, Saturated design, S-optimal design

A new class of row-column designs is proposed. These designs are saturated in terms of eliminating two-way heterogeneity with an additive model. The proposed designs are treatment-connected, i.e., all paired comparisons of treatments in the designs are estimable in spite of the existence of row and column effects. The connectedness of the designs is justified from two perspectives: linear model and contrast estimability. Comparisons with other designs are studied in terms of A-, D-, E-efficiencies as well as design balance.

Exact D-Optimal Designs For ScheffÈ Quadratic Models With Mixtures

✦Mong-Na Lo Huang, National Sun Yat-sen University, No. 70 Lienhai Rd., Kaohsiung, International 80424 Taiwan, R.O.C., *lomn@math.nsysu.edu.tw*; Hsiang-Ling Hsu, National Sun Yat-sen University; Shian-Chung Wu, National Sun Yat-sen University

Key Words: Equivalence theorem, orthogonal polynomials, geometric-arithmetic inequality for matrices

In this work, exact D-optimal designs for ScheffÈ quadratic models with mixtures are investigated. In the mixture experiment considered, it is assumed the measure response depends only on the proportions of the q ingredients present in the mixture. By using the equivalence theorem for the approximate D-optimal designs and geometric-arithmetic inequalities for matrices, we show that for given sample size N large enough, there are exact D-optimal designs supported as evenly as possible on the supports of the approximate D-optimal designs for some q.

I-Optimal Designs For Mixture Experiments With Linear Inequality Constraints

✦Laura Lancaster, SAS Institute, 600 Research Drive, Cary, NC 27513, *Laura.Lancaster@jmp.com*; Christopher Gotwalt, SAS Institute

Key Words: Design of Experiments, I-Optimality, Mixture Experiments, Nonlinear Programming

Predictive capability is an important objective of many mixture experiments. Although I-Optimal designs would be useful in a mixture context because they minimize the average prediction variance over the design region, they have not been investigated. This is because creating I-Optimal designs for mixture experiments with linear inequality constraints on the factors can be challenging in two regards. First, the objective function involves a moments matrix whose calculation requires several integrals over a linearly constrained region. The other challenge involves finding the optimal design in the presence of linear inequality constraints, as nonlinear programming methods are needed for the optimization. We will show how a Monte Carlo method for generating random uniform points over an arbitrary linearly constrained region can be used to calculate the moments matrix and how the Wolfe reduced gradient method can be used for optimizing the design. In addition, we demonstrate the improved predictive capacity of I-Optimal mixture designs by comparing them with other methods for generating mixture designs.

Comparing Designs For One-Step Response Surface Methodology

✦David Edwards, Virginia Commonwealth University, 1015 Floyd Avenue, P.O. Box 843083, Richmond, VA 23284, *dedwards7@vcu.edu*

Key Words: factor sparsity, minimal aliasing design, one-step RSM, orthogonal array, screening

The sequential design approach to response surface exploration is often viewed as advantageous as it provides the opportunity to learn from each successive experiment with the ultimate goal of determining optimum operating conditions for the system or process under study. Recent literature has explored factor screening and response surface optimization using only one three-level design to handle situations where conducting multiple experiments is prohibitive. The most straightforward and accessible analysis strategy for such designs is to first perform a main-effects only analysis to screen important factors before projecting the design onto these factors to conduct response surface exploration. This article proposes the use of optimal designs with minimal aliasing (MA designs) and demonstrates that they are more effective at screening important factors than the existing designs recommended for single-

design response surface exploration. For comparison purposes, we construct 27-run MA designs with up to 13 factors and demonstrate their utility using established design criterion and a simulation study.

A Theory of General Minimum Lower-Order Confounding for Factorial Designs

✦Runchu Zhang, Nankai University and Northeast Normal University, School of Mathematical Sciences, Nankai University, Tianjin, 300071 China, *zhrch@nankai.edu.cn*

Key Words: Clear effect, Confounding, Blocked design, Factorial design, Minimum aberration, Split-plot design

This talk presents a theory of general minimum lower order confounding (GMC) for fractional factorial designs proposed by Zhang, Li, Zhao and Ai(2008) and its recent developments. A characterization of GMC criterion to s-level case was given; Some essential properties of existing criteria (MA, MEC and CE) were discovered by GMC theory; All the 2^{n-m} GMC designs with $N/4+1 \le n\le N-1$ have been obtained, where N is run number and n factor number. The GMC criteria for blocked factorial designs were established, including B-GMC and B^1-GMC for single block variable case and B^2-GMC for multi block variable case; All the B^1-GMC $2^{n-m}:2^r$ designs with $5N/16+1\le n\le N-1$ have been obtained, where r is block factor number; The GMC theory was extended to the case of split-plot designs; The factor aliased effect-umber pattern (F-AENP) was introduced for arranging factors. The works are included in ZLZA(2008), Zhang and Mukerjee(2009a,b), Hu and Zhang(2010), Li,Zhao and Zhang(2010), Zhang and Cheng(2010), Cheng and Zhang(2010),Zhao et al(2010), Zhang,Li and Wei(2010), Wei et al(2010),Zhang,Wei and Li(2010), and Zhou,Balakrishnan and Zhang(2010).

Achieving Covariate Balance In Factorial Designs Using The Finite Selection Model

✦Li Zhu, Harvard University, Department of Statistics, 7th floor, 1 Oxford St, Cambridge, MA 02138, *lizhu@fas.harvard.edu*; Tirthankar Dasgupta, Harvard University, Department of Statistics

Key Words: Factorial Designs, Covariates, Finite Selection Model, Assignment Mechanism, Simulations

Factorial designs are widely used in agriculture, engineering and social sciences to study causal effects of several factors simultaneously on a response. Problems of application of factorial designs in social and medical experiments is the potential confounding of factorial effects with effects of a large number of covariates that are usually associated with the experimental units. Homogeneous groups of experimental units is therefore a challenge. In this research we explore an application of the Finite Selection Model (FSM) with the objective of allocating experimental units to different treatment groups so that the covariates are reasonably balanced across all factorial effects of interest. Theoretical properties of the design are established. Extensive simulation has demonstrated advantages of FSM method over traditional methods.

129 Small Area Estimation: Applications to Health Data

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

Monday, August 1, 8:30 a.m.–10:20 a.m.

Modeling Health Insurance Coverage Estimates For Minnesota Counties

✦Joanna Turner, University of Minnesota, 2221 University Ave. SE, Ste. 345, Minneapolis, MN 55414, *turn0053@umn.edu*; Peter Graven, University of Minnesota - SHADAC

Key Words: health insurance coverage, small area estimates, MNHA, county estimates, ACS

The University of Minnesota's State Health Access Data Assistance Center (SHADAC) is investigating the feasibility of producing model-based uninsurance estimates for Minnesota counties. County level estimates of health insurance coverage are frequently requested data from the Minnesota Health Access Survey (MNHA), but the survey sample size is not large enough to support estimates for this small geographic area. We build on the work of producing small area estimates for Oklahoma (SHADAC 2009) and use research findings from the Small Area Health Insurance Estimates program to develop a methodology using Bayesian techniques to model Minnesota county level estimates of uninsurance. We explore using the American Community Survey, the MNHA, administrative data sources such as unemployment claims and Medicaid/MHCP enrollment data and correlations over time and geography of these sources for the model. We develop and test the model on the 2009 MNHA data with plans to implement with the 2011 data.

Extending The Use Of A Local Health Survey To Finer Spatial Resolutions

✦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: small area estimates, statistical practice, health surveys, public health, monitoring and evaluation, built environment

The New York City (NYC) Community Health Survey (CHS) is a random-digit dial telephone survey of non-institutionalized adults conducted annually since 2002 by NYC's Department of Health and Mental Hygiene (DOHMH). The CHS collects information describing a wide variety of health outcomes and behaviors and is extensively used in planning and evaluation. The survey is designed to obtain estimates at the neighborhood-scale, defined using respondent zip codes and provides direct estimates for 34 neighborhoods (from 180 populated zip codes). Due to growing interest in linking the CHS to built environment and environmental exposure data the DOHMH has implemented procedures to extend the use of the survey to the zip code scale. We use unit-level small area estimation techniques calibrated to officially-released neighborhood estimates. Despite such methods being well developed in the statistical literature, we encountered a number of statistical challenges including: handling observations not associated

with zip codes, collapsing small zip codes, adjusting and accounting for complex survey design, choice of population controls, and estimating trends consistent with official estimates.

Application Of Hierarchical Bayesian Model With Poststratification For Small Area Estimation From Complex Survey Data

✦Vladislav Beresovsky, National Center for Health Statistic, 3311 Toledo Rd Room# 3226, Hyattsville, MD 20782, *vberesovsky@cdc.gov*; Catharine W Burt, National Center for Health Statistics; Van Parsons, National Center for Health Statistics; Nathaniel Schenker, National Center for Health Statistics; Ryan Mutter, Agency for Healthcare Research and Quality

Key Words: health care utilization, small area estimation, multilevel logistic regression, poststratification, hierarchical Bayesian model

Small area estimation from stratified multilevel surveys is well known to be challenging because of extreme variability of survey weights and the high level of data clustering. These challenges complicate the use of surveys such as the National Hospital Ambulatory Medical Care Survey to produce county- and state- level estimates of health care indicators. We focus on two indicators, the proportions of emergency department (ED) visits with asthma diagnoses and with injury diagnoses. We used multilevel logistic regression models to predict the proportion of visits of interest conditional on poststratification cells defined by independent variables and stratification information. Hospital level information including counts of total ED visits and county level demographic, social, economic and health care variables were obtained from the Verispan Hospital Database and Area Resource File. Combining these multiple sources of information increases the efficiency of predicted proportions. We evaluated models by comparing predictions with estimates based on administrative data from the Healthcare Cost and Utilization Project (HCUP) databases.

A Three-Part Model For Survey Estimates Of Proportions

✦Samuel Szelepka, U.S. Census Bureau, 2605 Fort Farnsworth Rd, Alexandria, VA 22303, *samuel.szelepka@census.gov*; Mark Bauder, U.S. Census Bureau

Key Words: SAHIE, Health Insurance, Small Area Estimation

In the Small Area Health Insurance Estimates program, we produce model based estimates of health insurance coverage for demographic groups within states and counties. In part of the model, we model survey estimates of proportions insured, conditional on the actual proportions. These survey estimates are bounded between zero and one, and have positive probabilities of being exactly zero and exactly one. Thus, assuming normality, or any continuous distribution is questionable. Many survey estimates are one, and some are zero because of high proportions insured and small sample sizes. To handle the boundedness and probability masses at zero and one, we have developed a "three-part" model. We model the probability that a survey estimate is zero, the probability that it is one, and its distribution conditional on not being zero or one. The models for probabilities of zero and one depend on the actual proportion, sample size, and parameters that are estimated. Conditional on not being zero or one, we assume a beta distribution. In this paper, we describe the three-part model, and present results from using the model and diagnostics of model fit.

Methods And Results For Small Area Estimation Using Smoking Data From The 2008 National Health Interview Survey

✦Neung Soo Ha, National Center for Health Statistics, Metro IV Building, 3311 Toledo Road, Hyattsville, MD 20782, *jvz5@cdc.gov*; Van Parsons, National Center for Health Statistics; Partha Lahiri, University of Maryland at College Park

Key Words: National Health Interview Survey, Small area estimation, Fay-Herriot method, Variance estimation, Bench-marking, Hierarchical modeling

The NHIS is designed to provide national level and regional estimates for health related conditions; however, it is not designed to produce estimates at the state or county level. We propose some small area estimation methods to provide state and sub-state level domain estimates and illustrate such methods using smoking data from the 2008 NHIS with the focus on area-level modeling. Since its introduction, the Fay-Herriot model (FH) has been a widely used SAE method to produce small area estimates from larger surveys. We apply a Bayesian MCMC approach for estimation via a posterior distribution. Practical survey considerations and needs require methodology adaptations. First, in practice the area level variances must be estimated. Traditional design-based estimators become unstable when the sample size within an area is small. We will investigate estimation methods that lead to more stable variance estimators. Second, aggregations of the small area estimates are preferred to be close to the original design-based total domain estimate. We will use the FH model along with a bench-marking technique to further improve the accuracy of the estimates and satisfy this preference.

The Development Of State Estimates From A National Health Survey

✦Wendy Van de Kerckhove, Westat, 1600 Research Blvd, Rockville, MD 20850, *wendyvandekerckhove@westat.com*; Robert E. Fay , Westat; Leyla K. Mohadjer, Westat; Lester R. Curtin, National Center for Health Statistics

Key Words: Small area estimation, NHANES, CHIS

The National Health and Nutrition Examination Survey (NHANES) is an on-going survey of the United States population that collects health and nutrition information through medical examinations. The support of the sampled communities is beneficial in obtaining cooperation from sampled persons and setting up mobile examination centers in the area. To provide valuable information to the communities and extend the usefulness of the NHANES data, NCHS, the survey's sponsor, decided to support research to develop local area estimates. This paper describes the method used to produce estimates for the state of California. The method needed to address the challenges of producing state estimates from a national survey, combining multiple years of data from different sample designs, and allowing estimation of multiple characteristics of interest. To produce estimates for multiple characteristics, we developed an approach to produce weights for NHANES sample cases in CA during the period 1999-2006. We report on a quasi-design-based approach we developed to combine data from the non-self-representing PSUs that had been sampled under two different stratification designs.

State And Local Wireless And Landline Estimates

✦Nadarajasundaram Ganesh, NORC at the University of Chicago, 4350 East-West Highway, 8th Floor, Bethesda, MD 20814, *ganesh-nada@norc.org*; Michael Davern, NORC at the University of Chicago; Stephen Blumberg, National Center for Health Statistics, Centers for Disease Control and Prevention ; Julian Luke, National Center for Health Statistics, Centers for Disease Control and Prevention ; Miche Boudreaux, University of Minnesota; Karen Soderberg, SHADAC, University of Minnesota

Key Words: NHIS, Small Area Estimation, State Wireless Prevalence Estimates

High-quality state-level prevalence estimates of wireless-only, wireless-mostly, landline-mostly, landline-only and non-telephone households are essential when weighting data from random-digit-dial telephone surveys conducted at the state level. Currently these estimates are available from the National Health Interview Survey (NHIS) at only the national and regional level. We will present recent state and local model-based wireless and landline estimates, produced using a combination of January 2007-June 2010 NHIS data, 2006-2009 American Community Survey data, and 2007-2010 data on listed households from infoUSA.com. We discuss our methods and observed efficiency gains using our modeling approach, and we present graphs/choropleth maps indicating changes over time. Finally, we discuss limitations of our modeling approach and also possible future improvements.

130 Issues with Adaptive Designs and DMCS ■

Biopharmaceutical Section

Monday, August 1, 8:30 a.m.–10:20 a.m.

Sample Size Reassessment In Fixed Duration Survival Trials

✦Alison Pedley, Merck, Boston University, , *alisonpedley@gmail.com*; Ralph B. D'Agostino, Sr, , Boston University ; Michael Pencina, Boston University; Joseph Massaro, Boston University

Key Words: Conditional Power, Interim Analysis, Sample Size Re-estimation, Survival Trials

The log-rank test is often performed in randomized clinical trials designed to assess the superiority of an experimental treatment over a control with respect to a time to event endpoint. Due to uncertainties in the design stage and the desire to accelerate drug development while reducing costs, adaptive designs allowing for sample size adjustment at the time of interim analysis based on the conditional power of observing a significant result by the end of the trial are becoming increasingly popular. In this presentation, methodology for a 2-stage adaptive design based on the log-rank test is developed in the setting of fixed duration trials by adapting methodology originally proposed by Li et al (2002, 2005) for maximum information trials. Simulations were performed to evaluate the performance of the new methodology. By redefining the relationship between the observed number of events and the final critical value and removing the restriction on the maximum value of the interim efficacy boundary, the methodology developed achieved the desired interim conditional power while still maintaining control of the type I error rate.

A Bayesian Adaptive Dropping-Arm Design Using Multiple Interim Evaluations For Decision Making

✦Yili L. Pritchett, Abbott Laboratories, 100 Abbott Park Road, Abbott Park, IL 60064 USA, *yili.pritchett@abbott.com*; Shihua Wen, Abbott Laboratories

Key Words: Adaptive dropping-arm design, Bayesian predictive probability, Futility criteria

Adaptive dropping-arm design allows studies to initially evaluate multiple doses with equal sample size, and then drop "losers" from further enrollment by applying pre-specified criteria to accumulated interim data in the middle course of the study. In most dropping-arm designs, the decision is based on the results of a single interim evaluation, which requires the right choice of interim timing and consistence of trial data to make the right decision. While easy to implement, the decision made at interim in such design is prone to mistake due to possible errors included in interim data or heterogeneity of data between those collected at early and late stage of the study. In this presentation, we introduce a dropping-arm design where futility criteria are made based on the pattern of consecutively calculated Bayesian predictive probabilities at multiple interim evaluations. The adaptive algorithm and decision rules will be described. Simulation details and the design operating characteristics will also be presented.

Simulation And Optimization Of An Adaptive Phase 2 Dose-Finding Study

✦Yannis Jemiai, Cytel, Cambridge, MA 02139 USA, *yannis@cytel.com*

Key Words: adaptive, dose-finding, phase 2, optimization, simulation

The proper design of phase 2 studies in drug development is critical in appropriately assessing the safety profile and collecting preliminary evidence of efficacy with which to move forward. A case study on the planning of a response-adaptive dose-finding study in neuropathic pain will be presented, including statistical and operational aspects of the trial design. Consultation with an inter-disciplinary group of professionals allowed an integrated evaluation of the potential implementation hurdles and benefits of the adaptive design. Intense simulation studies were required to produce the necessary evaluations under a variety of possible scenarios and design alternatives. Trial simulation software was particularly critical for the planning exercise.

Tracking Emerging Infectious Disease Epidemics In Real-Time Using Spectral Bayesian Data Assimilation

✦Ashok Krishnamurthy, University of Colorado Denver, 1175 Albion St Apt 409, Denver, CO 80220 United States, *ashokkrish@gmail.com*; Jonathan Beezley, University of Colorado Denver; Loren Cobb, University of Colorado Denver; Jan Mandel, University of Colorado Denver

Key Words: spatial epidemiology, data assimilation, disease tracking, bayesian design, traveling wave, FFT

We apply a discretized stochastic version of the spatial susceptible-infectious-removed (S-I-R) compartmental model of epidemiology to track in real time the incidence on new cases of a fast moving simulated traveling wave of an infectious disease epidemic. We use spectral approximation of the covariance by FFT and wavelets to obtain fast methods that can handle spatial simulations with large state vectors on a laptop in real time. These improved tracking methods derive their speed gains from smaller required ensemble sizes and more efficient algebra. Finally, we extend our methods to incorporate long-distance transportation, in order to track the rapid geographical spread of infectious diseases in human populations.

Futility Rule And Unblind Sample Size Re-Estiamtion In Phase Iii Drug Development: Method And Practical Example

✦Kyoungah See, Lilly, US Lilly Cooporate Center, Drop Code 1538, Indianapolis, IN 46285 USA, *seeky@lilly.com*

Key Words: Conditional Pronababilities, Operating chracteristics, Stopping Rules

In this talk we shall focus our discussion on futility monitoring and sample size re-estimation in confirmatory phase 3 clinical trials based on unblinded interim results through practical case and simulation studies. These techniques are useful when a lack of historical information and uncertainty about variability in the data and primary effect size exist. In such settings, it is rather essential to unblind the interim data and an independent interim analysis committee is required. We shall discuss practical examples including some statistical methodology and some operational and regulatory issues. Sample size re-estimation allows an adjustment in sample size during the trial. Starting with a small but reasonable up-front sample size commitment, an increase is applied only if interim results meet the prespcified rule. The sample size re-estimation method follows Proschan et al.(2006), Gao et al. (2008), and Mehta and Pocock (2010). Analyses based on binary endpoints are explored.

131 Topics with Non-inferiority and Eqiuvalence Trials

Biopharmaceutical Section

Monday, August 1, 8:30 a.m.–10:20 a.m.

Non-Inferiority Margin For Longitudinal Data, The Timolol Ophthalmic Solution Example

✦Rima Izem, Food and Drug Administration/CDER, Silver Spring, MD 20993 US, *rima.izem@fda.hhs.gov*

Key Words: Non-inferiority, longitudinal data, ophthalmic drug

In this presentation we share our derivation of an ophthalmic drug's non-inferiority margin for a longitudinal primary endpoint using limited historical data. In addition to the derivation, we briefly discuss non-inferiority design and testing strategies for longitudinal endpoints assessing both the short term treatment effect and the long term maintenance of treatment effect. We use the example of Timolol Ophthalmic solution 0.5% twice a day (TIM) for the indication of lowering elevated intra-ocular pressure in subjects with open angle glaucoma or ocular hypertension. TIM is one of the recommended and the most used active control comparator in Phase 3 trials for this indication. Deriving a non-inferiority margin for TIM is difficult because there is limited longitudinal data comparing TIM to placebo. We show how to account for diurnal variation and effect of the drug over time in our non-inferiority derivation.

A More Powerful Test Based On Ratio Distribution For Retention Non-Inferiority Hypothesis

✦Ling Deng, Clinical Biostatistics, Johnson & Johnson Pharmaceutical R&D, 920 Rt. 202, Raritan,, NJ 08869, *ldeng6@its.jnj.com*; Gang Chen, Johnson & Johnson

Key Words: non-inferiority trial, fraction retention, ratio test, ratio distribution, synthesis method

Rothmann et al. (Stat. Med. 2003; 22:239-264) proposed a method for the statistical inference of fraction retention non-inferiority (NI) hypothesis. A fraction retention hypothesis is defined as a ratio of the new treatment effect verse the control effect. One of the major concerns using this method is that with a limited sample size, the power of the study is usually very low. To improve power, Wang et al. (J. Biopharma. Stat. 2006; 16:151-164) proposed a ratio test based on asymptotic normality theory with a strong assumption of equal variance of the NI test statistic under null and alternative hypotheses in a sample size calculation. However, in practice, such assumption is generally questionable. This assumption is removed in the ratio test proposed in this paper, which is derived directly from a Cauchy-like ratio distribution. In addition, using this method, the fundamental assumption used in Rothmann's test, that the observed control effect is always positive, is no longer necessary. Without assuming equal variance under null and alternative hypotheses, the sample size can be significantly reduced if using the proposed ratio test for a fraction retention NI hypothesis.

Justification Of Non-Inferiority Design In Infection Disease Studies

✦Chunzhang Wu, Astellas Pharma Global Development, Inc., 3 Parkway North, Deerfield, IL 60015, *chunzhang.wu@us.astellas.com*

Key Words: non-inferiority margin, sensitivity, placebo effects, random effects

In infection disease area, non-inferiority (NI) design is a most acceptable methodology to design the clinical trials to support regulatory submission in recent decades. However the justification of the design, in particular the sensitivity and the selection of non-inferiority margin, is a critical challenge for a successful clinical trial. Using some successful historical and existing clinical trials related to antifungal disease, this paper illustrates statistical/clinical justification of NI design in accordance with major regulatory guidance. The sensitivity and the selection of non-inferiority margin is a primary focus. It is also in depth discussed that how to obtain conservative and reliable estimate of placebo effect when placebo control trials were not feasible in infection disease area in most health care practice.

Statistical Considerations In Assessing Assay Sensitivity In A Non-Inferiority Trial

✦Isaac Nuamah, Johnson & Johnson PRD, 1125 Trenton-Harbourton Road, Titusville, NJ 08560, *inuamah@its.jnj.com*

Key Words: non-inferiority, assay sensitivity, gold standard design

In clinical trials where non-inferiority of a new experimental drug to an active control has to be shown, it has been suggested to include a placebo arm (if ethically justifiable) in order to have an internal evidence to demonstrate assay sensitivity. In such a three-arm trial (sometimes referred to as 'gold standard design'), testing for non-inferiority follows a two-step hierarchical process. In the first step, superiority of either the experimental or active control to placebo is performed, and in the second step, non-inferiority is assessed. For an active control non-inferiority study (without a concurrent placebo arm), assay sensitivity cannot be directly assessed in that trial and has to depend on appropriate trial conduct and historical evidence of sensitivity of active drug effect. In this talk, we discuss the statistical issues needed to demonstrate assay sensitivity when (1) there is a placebo arm in that trial, the so-called 'gold standard design' and (2) when the only comparator is the active control arm. Data from a psychiatry clinical trial are used to illustrate these issues.

Non-Inferiority Trial Designs Lack Of Historical Study Results

✦Guozhi Gao, Sanofi-aventis, 02140 MA, *guozhigao@gmail.com*

Key Words: non-inferiority

In clinical trials non-inferiority trial designs have gained popularity in various therapeutic areas. These designs often rely on effects of active control (e.g. standard of care) over placebo that are available from historical study results, usually by pooling treatment effects from multiple historical studies. However, not always this information is available. In this talk we focus on a practical problem where the standard of care (S) plus intervention B is a popular treatment option for a certain disease population, and the objective of the trial is to test the null hypothesis that the combo therapy: experimental treatment A + intervention B is non-inferior to S + B. The challenge of this problem is that the effect of S + B over placebo is unclear because the added intervention B was due to advance of science (or increased knowledge of the disease). The lack of historical study results places a hurdle in the non-inferiority trial design, including choosing margins and calculating sample size. We will discuss approaches to deal with this problem.

132 Multiplicity and Multiple Comparisons (I) ■

Biopharmaceutical Section

Monday, August 1, 8:30 a.m.–10:20 a.m.

Comparison of Gatekeeping and Other Testing Methods for Identifying Superior Drug Combinations in Bi-Factorial Designs with Isotonic Parameters

✦Julia Soulakova, University of Nebraska-Lincoln, , *jsoulakova2@unl.edu*

Key Words: closure principle, family-wise error rate, maximum test, multiple testing

Several multiple testing procedures are discussed with respect to the problem of detecting combination drug superiority for a bi-factorial design with monotone gains. The testing methods include the generalized maximum test procedure (GMAXP) proposed by Soulakova (2009), and several gatekeeping strategies: the general multistage gatekeeping method (GMGP) proposed by Dmitrienko, Tamhane and Wiens (2008), and the serial Bonferroni gatekeeping method (TREE), proposed by Dmitrienko, Wiens, Tamhane and Wang (2007). The GMGP with the truncated Holm and Hochberg components is discussed. In addition, the truncated Sidak-Holm component is proposed. It is shown by simulations that the GMAXP achieves higher power if there are relatively many superior combinations, while the gatekeeping methods perform better if there is a single or just a few superior combinations. While in some cases the GMGP with the truncated Sidak-Holm component outperforms the GMGP with the truncated Hochberg component, the observed power advantages are not substantial. Thus, the GMGP with the truncated Hochberg component is recommended as it can dominate the other methods when the Min test statistics are independent.

Asymptotically Efficient Sequential Tests Of Multiple Hypotheses

✦Shyamal Krishna De, University of Texas at Dallas, Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75080, *skd081000@utdallas.edu*; Michael Baron, University of Texas at Dallas

Key Words: multiple comparisons, stopping rule, asymptotic optimality, Pitman alternative, sequential probability ratio test, familywise error rate

A number of sequential experiments include multiple statistical inferences such as testing multiple hypotheses, constructing simultaneous confidence sets, or making other decisions involving multiple parameters or multiple measurements. Examples include sequential clinical trials for testing both safety and efficacy of a treatment, quality control charts monitoring a number of measures, acceptance sampling requiring several criteria, and so on. In each application, it is essential to get a result of each individual inference instead of combining results into one procedure giving one global answer. A sequential procedure for testing multiple hypotheses is presented that achieves the asymptotically optimal rate of the expected sample size under the strong control for Type I and Type II familywise error rates. Stopping rules for the sequential testing are proposed, and the form of asymptotically optimal stopping boundaries is derived under Pitman alternatives. Stopping rules are compared; the resulting cost saving and reduction of error rates are analyzed.

Multiplicity Adjustment For Multiple Treatments (Doses) And Multiple Endpoints

✦Nancy Ying Liu, Merck & Co., Inc., 126 E. Lincoln Ave, P.O.Box 2000, Rahway, NJ, NJ 07065, *nancy_liu@merck.com*; Jing Li, Merck & Co., Inc.; Ziliang Li, Merck & Co., Inc.; Amarjot Kaur, Merck

Key Words: multiple doses, multiple endpoints

In order to better characterize the dose responses and to speed up the drug development process, some sponsors choose to combine the Phase IIb dose-ranging study with the Phase III study (as a Phase IIb/III study) to form the basis of future submission. As a result, more complicated multiple testing problems arise related to the multiple doses and multiple endpoints. This presentation compares several multiplicity adjustments as illustrated by a phase IIb/III case study. Several multiplicity adjustments with strong control or weak control of the familywise error rate (FWER) are reviewed and illustrated by the case study. Strategies considered with strong control of the FWER include the serial and the parallel gatekeeping procedures, the gatekeeping procedure based on the Dunnett test, and the truncated Hochberg multiple test procedure (MTP); several strategies with weak control of the FWER, such as parallel step-down testing procedure and Hochberg testing procedure, are also considered. Property of each procedure is compared using simulations. Some recommendations on choosing multiplicity strategy for phase IIb/III clinical studies are provided at the end of the presentation.

Development Of Multiple Testing Strategies In Confirmatory Clinical Trials

✦Brian A Millen, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, *bmillen@lilly.com*; Alex Dmitrienko, Eli Lilly and Company

Key Words: gatekeeping, familywise error rate

This paper discusses practical considerations in the development of multiple testing strategies in confirmatory clinical trials. General principles for development of testing procedures are introduced. Clinical trial examples are provided to illustrate these general principles and other key concepts. Emphasis is placed on incorporating trial-specific information in the development and selection of the multiple testing procedure to meet desired performance criteria.

Multiple Comparisons For Partial Covariance Matrices Of Two Treatment Groups In Clinical Trial

✦Yoshiomi Nakazuru, Tokyo University of Science, Tokyo, Japan, *yoshiomi.nakazuru@pfizer.com*; Takashi Seo, Tokyo University of Science

Key Words: Multiple comparisons, Covariance matrix, Clinical trial

Generally, the main focus of clinical trial is to demonstrate the efficacy of new treatment in terms of mean(s) of the primary endpoint(s). In some cases, however, variance(s) of the endpoints also may be the focus of interest. For example, a treatment with smaller variability, and thus more predictable efficacy, may be preferable, given two treatments with equal efficacy in terms of the means. In this presentation we consider multiple comparison procedure for partial covariance matrices of endpoints in clinical trials to demonstrate the superiority of a new treatment in terms of the variability. First, we review and discuss the tests for the equality of two covariance matrices based on the Union-Intersection test procedure. Second, we propose a multiple comparison procedure for partial covariance matrices that limiting the number of comparisons and compare its performance with the method discussed in first section, using Monte Carlo simulation with a normality assumption. The simulation results suggest that powers of the proposed procedure are generally higher than those of the previous method, while keeping type I error rates nearly within the nominal level.

Multiplicity Issues In Clinical Trials With Co-Primary Endpoints, Secondary Endpoints And Multiple Dose Comparisons

✦Haiyan Xu, Johnson & Johnson PRD, NJ , *hxu22@its.jnj.com*; Pilar Lim, Johnson & Johnson PRD

Key Words: Multiple endpoints, Multiplicity, Clinical trial, Gatekeeping

In clinical trials there are situations when the overall type I error rate needs to be controlled across co-primary endpoints and secondary endpoints. For example, a health authority may require improvement in both pain and functioning as compared to placebo for the treatment of pain due to osteoarthritis. The sponsor may also be interested in claiming efficacy in additional secondary endpoints in such a trial. The multiplicity issue can be further complicated by multiple dose comparisons. This paper proposes several methods that control the overall type I error for these situations. These methods are constructed using the closed testing principle and the partitioning principle. This paper also compares these methods with the IUT (intersection-union test) based method that is commonly used in testing co-primary endpoints.

Multistage Parallel Gatekeeping With Retesting

✦George Kordzakhia, Food and Drug Administration, 10903 New Hampshire Ave., Building 21, Room 4603, Silver Spring,20993-0002, MD 20993-0002 USA, *George.Kordzakhia@fda.hhs.gov*; Alex Dmitrienko, Eli Lilly and Company

Key Words: Multiple Comparisons, Familywise error rate, Parallel gatekeeping, Cosure principle, Mixture procedure

This talk introduces a general method for constructing multistage parallel gatekeeping procedures with a retesting option. This approach serves as an extension of general multistage parallel gatekeeping procedures (Dmitrienko, Tamhane and Wiens, 2008) and parallel gatekeeping procedures with retesting for two-family problems (Dmitrienko, Kordzakhia, and Tamhane, 2011). It was shown in the latter paper that power of parallel gatekeeping procedures can be improved by adding an additional retesting stage which enables retesting of the primary null hypotheses using a more powerful component procedure than the original procedure if all secondary null hypotheses are rejected. The new method enables clinical trial researchers to construct a class of more general multistage parallel gatekeeping procedures with retesting. The new procedures support multiple retesting of the primary and secondary families and do not require all secondary null hypotheses be rejected.

133 Next Generation Genetics: Hopes & Hypes ■●

ENAR, Scientific and Public Affairs Advisory Committee

Monday, August 1, 8:30 a.m.–10:20 a.m.

Genome-Wide Eqtl Mapping By Iterative Multivariate Adaptive Lasso (Imal)

✦Ting-Huei Chen, Department of Biostatistics, University of North Carolina, Chapel Hill, 2701 Homestead Road, APT 603, Chapel Hill, NC 27516, *tchen@bios.unc.edu*; Wei Sun, University of North Carolina, Chapel Hill; Fred Andrew Wright, Univ North Carolina

Key Words: Multivariate penalized regression, iterative adaptive Lasso, multiple loci mapping, gene expression QTL, variable selection

Genome-wide eQTL (gene expression quantitative trait loci) mapping aims to select important genetic markers that explain the variances of the gene expressions. Therefore, it can be treated as a variable selection problem with multiple responses and covariates. Many methods have been proposed to analyze each gene expression trait separately. However, it is well known that several genes could be co-regulated and linked to some common genetic markers. The trait-by-trait mapping strategy fails to take into account the possible correlation between traits, which may lead to a loss of power. We propose IMAL, a multivariate penalized regression method, for eQTL mapping. IMAL employs two penalties to fulfill the parsimonious model assumption and accommodate possible genetic hotspots. Moreover, IMAL incorporates the correlation among traits and is applicable for the high dimension low sample size setting. Asymptotic studies show that the penalty function of IMAL has better capability to handle high dimensional problem than most popular penalties. Empirical results from both simulation and real data analysis confirm that IMAL has improved variable selection performance than existing methods.

Principal Interactions Analysis For Repeated Measures Data: Application To Gene-Gene, Gene-Environment Interactions

Bhramar Mukherjee, University of Michigan; ✦Yi-An Ko, University of Michigan, Department of Biostatistics, SPH, 1415 Washington Heights, Ann Arbor, MI 48109, *yianko@umich.edu*; Tyler J VanderWeele, Harvard University; Anindya Roy, University of Maryland Baltimore County; Sung Kyun Park, University of Michigan; Jinbo Chen, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine

Key Words: column interaction, epistasis, intraclass correlation, likelihood ratio test, non-additivity, Wishart matrix

Modeling gene-gene, gene-environment interactions with repeated measures data on a quantitative trait is considered. Classical models proposed by Tukey and Mandel using cell means of a two-way classification array are effective to detect interactions in presence of main effects, but they fail under misspecified interaction structures. We explore additive main effects and multiplicative interaction (AMMI) models, which are based on a singular value decomposition of the cell means residual matrix after fitting additive main effects. AMMI models provide summaries of subject-specific and time-varying contributions to the leading principal components of the interaction matrix and allow geometric representation of the structure. We call this analysis "Principal Interactions Analysis" (PIA). It is illustrated using data from a longitudinal cohort study. Simulation studies were carried out under classical and common epistasis models to reveal PIA properties in comparison with the classical alternatives. AMMI performs reasonably across a spectrum of interaction models. AMMI test, however, may not be very powerful for common epistasis models unless epistasis occurs without main effects.

A Statistical Procedure to Evaluate Agreement of Differential Expression for Translational Cross-Species Genomics

✦Cuilan Lani Gao, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, *cuilan.gao@stjude.org*; Stan Pounds, St. Jude Children's Research Hospital

Key Words: genomics, cross-species genomics, microarray, agreement, permutation

An important problem in translational genomics is to evaluate the fidelity of an animal model of a human disease. We developed the agreement of differential expression (AGDEX) procedure to evaluate the agreement of the results of a differential expression experiment using an animal model with those of a similar experiment using humans. AGDEX uses an agreement statistic to measure the similarity of expression differences for pre-defined sets of ortholog-matched genes. Significance is determined by permutation. In two cancer genomics studies, AGDEX was used to determine that a brain tumor in a mouse model showed a similar gene expression profile to that of a specific human brain tumor subtype. In both studies, these results were confirmed by subsequent laboratory investigation which revealed that the model tumor shows remarkable histological similarities to the identified human tumor. The combined results lead to the identification of the cell of origin for two different types of brain tumors. These examples provide compelling proof-of-principle that AGDEX can be a useful tool for biological discovery. Future research should further develop AGDEX and related concepts.

A Robust Model For Multilocus Population Genetics With Selfcrossing Rate

✦Jingyuan Liu, Department of Statistics, Penn State Univ., 325 Thomas Bldg, Penn State Univ., University Park, PA 16802, *jul221@psu.edu*

Key Words: Gametic linkage disequilibrium, zygotic linkage disequilibrium, Hardy-Weinberg equilibrium, non-equilibrium population, molecular marker

A fundamental assumption used for current multilocus analysis approaches is Hardy-Weinberg equilibrium. Given the fact that natural populations are rarely panmictic, these approaches will have a significant limitation for practical use. We present a robust model for multilocus linkage disequilibrium analysis which does not rely on the assumption of random mating. The new model capitalizes on Weir's definitions of zygotic disequilibria and is based on an open-pollinated design in which multiple maternal individuals and their half-sib families are sampled from a natural population, and it's applicable for both monoecious and dioecious plants. This design captures two levels of associations: one is at the upper level that describes the pattern of co-segregation between different loci in the parental population and the other is at the lower level that specifies the extent of co-transmission of homologous alleles at different loci from parents to their offspring. An MCMC method was implemented to estimate genetic parameters that define these associations. Simulation studies were used to validate the statistical behavior, and real data analysis was also provided.

Likelihood Ratio Test Process For Quantitative Trait Loci Detection

✦Charles-Elie Rabier, Department of Statistics and Department of Botany, University of Wisconsin, WI 53703, *rabier@stat.wisc.edu*; Jean-Marc AzaÔs, Institut de MathÈmatiques de Toulouse, UniversitÈ Paul Sabatier, France; CÈline Delmas, Station d'AmÈlioration GÈnÈtique des Animaux, INRA, France

Key Words: QTL Detection, Likelihood Ratio Test, Mixture models, Chi Square Process, Gaussian process, Interval Mapping

We address the problem of detecting Quantitative Trait Loci, so-called QTLs (genes influencing a quantitative trait which is able to be measured) on a given chromosome (modelized by a segment [0,T]). Lander and Botstein (1989) proposed the "Interval Mapping" : with the help of genetic markers, we scan the chromosome, performing a Likelihood Ratio Test (LRT) of the absence of a QTL at every location on [0,T]. So, it leads to a LRT process. In presence of several QTLs, Jansen (1993) and Zeng (1994) proposed the ``Composite Interval Mapping" (CIM), which consists in combining Interval Mapping on two flanking markers and multiple regression analysis on other markers. This way, the QTLs effects of the QTLs located outside the interval tested, are removed due to multiple regression analysis. In our work, we give the asymptotic distribution of the LRT process under the general alternative that there exist m QTL on [0,T]. It allows us to propose to estimate the number of QTLs and their positions using the LASSO. Our method does not require the choice of cofactors which is the main drawback of CIM. Using simulated data, we show that our method gives better performances than CIM.

Empirical Evaluation Of Methods For Identifying Recurrent Copy Number Variations Across Multiple Samples

✦Jeanette E Eckel-Passow, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, *eckel@mayo.edu*; Elizabeth J Atkinson, Mayo Clinic; Vernon S Pankratz, Mayo Clinic; Christopher G Scott, Mayo Clinic; Janet E Olson, Mayo Clinic; Julie M Cunningham, Mayo Clinic; Fergus J Couch, Mayo Clinic; Thomas Sellers, H. Lee Moffitt Cancer Center & Research Institute; Celine M Vachon, Mayo Clinic

Key Words: copy number, plink, composite, cover, cnvpack

Copy Number Variations (CNVs) are structural changes to regions of chromosomal DNA, resulting in a change in the normal diploid copy number. CNVs are detected for each sample independently and as a result of the inherent noise in the technology, most CNVs will contain sample-specific breakpoints. In order to perform downstream association analyses it is necessary to first define recurrent CNV regions across samples, specifically, to define a common chromosomal start and stop location across samples for each defined recurrent CNV. We empirically evaluated methods for identifying recurrent regions using the GENetic Epidemiology of MAMmographic Density (GENEMAM) study that evaluated 472 subjects from 90 families. The data were generated using the Illumina 660 SNP array. Sample-specific CNV regions were identified using PennCNV. Recurrent regions were identified using three methods: Plink and the COMPOSITE and COVER methods available in the R package cnvpack. We tested for an association between each recurrent insertion and deletion with breast density adjusting for familial correlation. The influence of each of the three methods on the association results will be presented.

Power And Sample Size Calculations For Snp Association Studies With Censored Time-To-Event Outcomes

✦Kouros Owzar, Duke University, , *kouros.owzar@duke.edu*; Zhiguo Li, Duke University; Sin-Ho Jung, Duke University

Key Words: SNP, survival, censoring, power, sample size, score test

For many clinical studies in cancer, germline DNA is prospectively collected for the purpose of identifying or validating Single Nucleotide Polymorphisms (SNP) associated with clinical outcomes. The primary clinical endpoint for many of these studies are time-to-event outcomes such as time of death or disease progression which are subject to censoring mechanisms. The Cox score test can be readily employed to test the association between a SNP and the outcome of interest. In addition to the effect and sample size, and censoring distribution, the power of the test will depend on the underlying genetic risk model and the distribution of the risk allele. We conduct a comprehensive review of the statistical power of the Cox score test under a variety of genetic risk models and risk allele distributions. Asymptotic formulas for power and sample size calculation will be presented.

134 Credit and Insurance Modeling ■●

Business and Economic Statistics Section, Scientific and Public Affairs Advisory Committee

Monday, August 1, 8:30 a.m.–10:20 a.m.

Cost-Sensitive Classification For Skewed Distribution: An Application To Credit Rating Model Validation

✦Raffaella Calabrese, University of Milano-Bicocca, ufficio 4012 Ed u7, Via Bicocca degli Arcimboldi 8, Milano, 20126 Italy, *raffaella.calabrese1@unimib.it*

Key Words: Cumulative Accuracy Profile, iso-performance line, classification error cost, discriminatory power index, optimal threshold

Receiver Operating Characteristic and Cumulative Accuracy Profile curves are used to assess the discriminatory power of credit rating models. Coherently with these curves, the Accuracy Ratio and the Area Under the Curve are discriminatory power indexes. To identify the optimal threshold on these curves, the iso-performance lines are used. These methodologies assume equal classification error costs, an unrealistic assumption in the application to credit risk. Another distinctive characteristic of the scoring model validation is the very low frequency of defaults. In order to take account of these characteristics, the curve of Classification Error Costs is proposed. Coherently with this curve, a discriminatory power index and a methodology to identify the optimal threshold are suggested. Finally, the methodological proposals are applied to data of Italian Small and Medium Enterprizes over the years 2004-2008.

Spatio-Temporal Modeling Of Credit Default Rates

✦Sathyanarayan Anand, Wharton School, Univ. of Pennsylvania, 3730 Walnut Street, Suite 400, Philadelphia, PA 19130, *sanand@wharton.upenn.edu*; Robert Stine, Wharton School, Univ. of Pennsylvania

Key Words: autoregressive, conditional, bayesian, spatial, temporal, mcmc

The availability of cheap computing power and pervasiveness of MCMC methods allow for the fitting of complex models that can account for intricate dependencies within data. We focus here on spatio-temporal models that have been widely studied with data on agricultural yields, disease mapping and climate. More specifically, we investigate the use of conditional autoregressive (CAR) models for fitting and predicting credit default rates in the US by county and by quarter. We begin with a hierarchical model that assumes default rates to be binomially distributed, but is computationally intractable, and derive a normal approximation that is validated and gives conjugate posterior distributions for efficient Gibbs sampling. We address issues of parameter identifiability when using CAR models along two dimensions, namely space and time. We draw inspiration from two-factor ANOVA models and reparameterize the CAR spatio-temporal effects in terms of contrasts, and derive prior distributions for the latter while maintaining the desired CAR structure. We compare the predictive performance of our model to other established methods on a full-scale simulated dataset and the credit default data.

Predicting Credit Application Fraud

✦Michiko I. Wolcott, Equifax, 1100 Abernathy Road Suite 300, Atlanta, GA 30328, *nolesan@equifax.com*

Key Words: Application Fraud, Fraud Exchange, Predictive Modeling, Credit

Credit application fraud modeling has received much attention in the recent years as another layer to risk mitigation at credit origination decisioning. While the legal framework under which fraud models and the associated data can be applied in practice varies widely, there is some fundamental learning that can be leveraged regardless of market, with some broad technical implications. This paper discusses some of the challenges and findings in application fraud modeling, addressing profiles of different types of application fraud, the ways in which different sources of data and different types of variables contribute to prediction of application fraud, how their contributions vary by the type of fraud predicted, and some technical challenges in application fraud prediction modeling, among others. As a case study, the development of a generic credit application fraud score in Canada is examined, in which the application information, data from a consortium fraud database, and credit bureau data are used among other data sources.

Learning Made Easy: A Marginalized Resample-Move Approach

✦Junye Li, ESSEC Business School, Singapore, 188064 Singapore, *li@essec.edu*

Key Words: State-space models, Particle filters, Parameter learning, State filtering, Resample-move, Stochastic volatility

Parameter learning in state-space models, especially in dynamic asset pricing models, is practically difficult. This paper proposes a simulation-based parameter learning method in the general state space models. First, the approach breaks up the interdependence of the hidden states and the static parameters by marginalizing out the states using a particle filter. Second, it proposes a Bayesian resample-move approach to this marginalized system. This marginalized resample-move is exact in the sense that for any fixed number of M particles used to in hidden states, it delivers sequential samples from the posterior distributions as the number of particles over the fixed parameters, N, goes to infinity. Simulation studies show that our learning method can deliver the same posterior outputs as standard MCMC methods, both in a linear Gaussian model and a nonlinear non-Gaussian one. More importantly, it provides posterior quantities necessary for full sequential inference and recursive model monitoring. Furthermore, the methodology is generic and needs little design effort. The algorithm is also implemented on real data for a stochastic volatility model and a credit risk model.

An Alternative Claim Amount Prediction For Deductible Insurance Policies

✦Heungsun (Sunny) Park, Hankuk University of Foreign Studies, Wangsan-ri Mohyun-myun, Yongin, International 449791 South Korea, *hspark@hufs.ac.kr*; Yongbum Jun, National Agricultural Cooperative Federation

Key Words: Best Predictor, Weighted Best Predictor, reinsurance

Best predictor is defined as the conditional expected value of a response. This paper proposes the modified version of best predictor with different weights for the insurance pricing market when considering deductibles. Insurance companies need to predict the claim amount for a policy or a group of policies with the least possible prediction errors while they don't have to take any risk for the claim below the deductible level. The same analogy can be applied to the reinsurance price making system.

Modeling And Evaluating Losses By Mixture Distribution

✦Min Deng, Maryville University at St. Louis, 650 Maryville University Drive, St. Louis, MO 63141, *mdeng@maryville.edu*

Key Words: Aggregate Claims, Counting distribution, Claim distribution, Mixture Distribution, Losses

This paper focuses on important aspects of modeling and evaluating losses. There is a lot of research on this topic is concentrated on Erlang distribution. In this paper we are going to discuss not only use Mixtures of ERlang Distribution as the tool to modeling and evaluating losses, and also use Mixture of other important distribution to modeling and evaluation losses. Some numerical examples to illustrate the results will be given.

Partially Adaptive Estimation of the Censored Regression Model

Randall A. Lewis, Yahoo! Research; ✦James B. McDonald, Brigham Young University, 153 FOB, Brigham Young University, Provo, UT 84602, *james_mcdonald@byu.edu*

Key Words: censored regression, Tobit, CLAD, SCLS, partially adaptive estimators

Data censoring causes ordinary least squares estimates of linear models to be biased and inconsistent. Tobit, nonparametric, and partially adaptive estimators have been considered as possible solutions. This paper proposes several new partially adaptive estimators that cover a wide range of distributional characteristics. A simulation study is used to investigate the estimators' relative efficiency in these settings. The partially adaptive censored regression estimators have little efficiency loss for censored normal errors and outperform Tobit and nonparametric estimators for non-normal distributions. An empirical example of out-of-pocket expenditures for a health insurance plan provides an example, which supports these results.

135 Financial Modeling

Business and Economic Statistics Section, Scientific and Public Affairs Advisory Committee

Monday, August 1, 8:30 a.m.–10:20 a.m.

Monitoring The Structural Change Of The Intraday Volatility Pattern

✦Robertas Gabrys, University of Southern California, Marshall School of Business, 3670 Trousdale Parkway, BRI 401O, Los Angeles, CA 90089-0809 USA, *robertas.gabrys@marshall.usc.edu*; Siegfried Hˆrmann, UniversitÈ Libre de Bruxelles; Piotr Kokoszka, Utah State University

Key Words: Change point detection, Intraday volatility, Functional data analysis, Sequential analysis

A functional time series consists of curves, typically one curve per day. The most important parameter of such a series is the mean curve. We propose two methods of detecting a change in the mean function of a functional time series. The change is detected on line, as new functional observations arrive. The general methodology is motivated by and applied to the detection of a change in the average intraday volatility pattern. The methodology is asymptotically justified by applying a new notion of weak dependence for functional time series. It is calibrated and validated by simulations based on real intraday volatility curves. We focus on the volatility of one minute returns on US stocks and indexes, but the statistical methodology we develop is very general, and can be applied to other asset classes, and to volume as well as volatility. In fact, it can be applied to any functional time series with a very general linear or nonlinear dependence structure, but we concentrate on detecting a change in the intraday volatility pattern.

Segmenting The Time Series Of A Market Index Using A Hidden Markov Model

✦Ziqian Tony Huang, Liautaud Graduate School of Business, University of Illinois at Chicago, 1865 N CRENSHAW CIR, VERNON HILLS, IL 60061, *ziqianhuang@hotmail.com*; Stanley L Sclove, Information & Decision Sciences Dept., Univ. of Illinois at Chicago

Key Words: time series, stock market index, S&P500, Bull and Bear, hidden Markov model

The time series of monthly rates of return (RORs) of the S&P 500 stock index for January, 1950, through September, 2010 (729 months, 728 RORs), was segmented by several methods, the up-down method, in which a month is declared to be a Bull month if the ROR was positive and a Bear month if it was negative, a Bull-Bear scoring method comparing one month's high, low and close to those of the preceding month, and some hidden Markov models (HMMs). HMMs with two and three states were fit and compared with a single distribution for the RORs. The single distribution appears to be non-Normal. State-conditional Normal distributions with different means and variances were fit for two and three states. The HMMs were scored by BIC. The model with two states was better. One state had a positive mean ROR, the other, a negative mean ROR. This corresponds to conventional notions of Bull and Bear states. The Bear state has a higher variance. Some comparison with ARIMA and ARCH models was made. Among these, there is evidence that an ARCH(3) model would be best, but not as good as the HMM model with two states.

Regime Switching in the Conditional Skewness of S&P 500 Returns

✦Mohammad Jahan-Parvar, East Carolina University, A 426 Brewster Building, Greenville, NC 27858, *jahanparvarm@ecu.edu*; Bruno Feunou, Duke University; Romeo Tedongap, Stockholm School of Economics

Key Words: GARCH, Regime Switching, Conditional Skewness, Volatility, Jumps, Financial Crisis

We introduce a simple regime switching GARCH model that captures features of the market returns in normal and crisis episodes better than the symmetric GARCH class, and is considerably easier to implement than Levy-driven models. Hamilton and Susmel (1994) introduce the idea of regime switching in the (G)ARCH literature. We borrow their concept, but apply it to the conditional skewness of returns. Conditional volatility of returns in our model follows a GARCH(1,1) process with Gaussian or Student-t errors, but during crisis episodes the volatility process switches to skewed GED GARCH. This switching happens when conditional skewness of the standardized residuals in the model deviate from near-zero values of a normal GARCH. In normal times, the conditional skewness of returns is practically zero. As a result, symmetric GARCH models are adequate for characterization of returns. On the other hand, during crisis periods, the conditional skewness may significantly deviate from zero. We use daily and monthly S&P500 excess returns for 1970-2010 period. We conduct extensive diagnostic testing for adequacy of the our model which show that it is as good or better than the alternatives.

The Structure And Estimation Of Discrete-State Time Series With Time-Varying Parameters

✦KAZUHIKO SHINKI, Wayne State University, 5055 Buckingham, Troy, MI 48098, *seikibunpu@gmail.com*

Key Words: time series, GARCH, heteroscedasticity, high frequency data, discrete-state, count data

High frequency financial time series is often discrete-state, since the price change between two transactions is typically zero, one or two ticks. Also GARCH-type time-varying parameters have to be included in the model when one wants to capture heteroscedasticity. While a few such models have been proposed since 90s, the probabilistic structures of the models are widely open. We clarify the structure and estimation of the autoregressive conditional multinomial (ACM) model

by Russell and Engle. Then we extend the model to (i) allow infinitely many possible states for evaluating tail-risk, and (ii) reduce the number of parameters.

Alpha Representation for Active Portfolio Management and High-Frequency Trading in Seemingly Efficient Markets

✦Godfrey Cadogan, Institute for Innovation and Technology Management, Ted Rogers School of Management, 575 Bay, Toronto, ON M5G 2C5 Canada, *godcent70@gmail.com*

Key Words: semi-martingales, stopped alpha process, market timing, efficient market

This paper's contribution to the market timing literature is twofold. First, it reconciles polemics between proponents of active portfolio management and advocates of ef?cient markets. Second, it introduces an empirical alpha process decomposed into (1) controlled jumps induced by news arrival, and (2) Brownian bridges over sojourn times for the alpha process. In particular, we use a canonical benchmark asset pricing model augmented with hedge factor mimicking derivatives for price discovery, and identification of asymptotic properties of alpha when portfolio strategy or investment style is unobservable. In a nutshell, the model predicts that the yin/yang of asset pricing anomalies are such that benchmark and hedge factor exposure(s) are semimartingales artifacts of an ef?cient market characterized by global martingales. Thus, we provide a solution to (Grossman and Stiglitz, 1980, pg. 395) open problems on the distribution of information among traders in seemingly ef?cient markets. In particular, informed traders have a "martingale system" while uninformed traders don't. Further, the local martingale processes driving factor exposures run on a different time clock than the underly

136 High Dimensional Covariance and Network Estimation

Section on Statistical Learning and Data Mining, Section on Statistical Computing

Monday, August 1, 8:30 a.m.–10:20 a.m.

Sure-Assisted Tapering Estimation Of Large Covariance Matrices

✦Feng Yi, University of Minnesota, 313 Ford Hall 224 Church St SE, Minneapolis, MN 55455, *yixxx064@umn.edu*; Hui Zou, University of Minnesota

Key Words: covariance matrix estimation, high-dimensional data, Frobenius norm, tapering estimation

A recent paper by Cai, Zhang and Zhou (2009) revealed the optimal minimax rates of convergence for high-dimensional covariance matrix estimation. Cai, Zhang and Zhou further proved that the tapering estimator can achieve the optimal rate if the tapering parameter is chosen according to the unknown sparse index parameter and the matrix norm. In order to apply the tapering estimator in practice, it is critically important to have a data-driven method to select the optimal tapering parameter. In this work we use Stein's unbiased risk estimation (SURE) to select the tapering parameter under the Frobenius norm. We show that with high probability SURE selects the optimal tapering parameter for the Frobenius norm without using the unknown sparse index parameter. For the matrix L_1 and L_2 norms, the SURE selected tapering parameter can be used to compute the optimal L_1 and L_2 tapering parameters. We conduct extensive simulation study to demonstrate the good performance of SURE-assisted Tapering estimator.

Doubly Shrinking Of Correlation Matrices

✦Sheng-Mao Chang, National Cheng Kung UNiversity, Department of Statistics, No. 1 Ta-Hsueh Road, Tainan, 70101 Taiwan, *smchang@mail.ncku.edu.tw*

Key Words: adaptive LASSO, correlation, generalized thresholding

Correlation matrices play an important role in may multivariate techniques. A good correlation estimation is therefore crucial in this kind of analysis. Sometimes, a correlation matrix is expected to be sparse due to the nature of the data or for the sake of simplification of interpretation. The generalized thresholding estimator possesses good properties such as sparsity, consistency and superior computational efficiency. However, the estimator is not always positive definite especially for not well-conditioned matrices. In this work, we propose a doubly shrinking method which shrinks tiny elements toward zero and then shrinks the correlation matrix toward the identity matrix. The performance of the proposed estimator is explored in terms of relative Frobenius norm. Theory and simulations were deduced under certain correlation matrices with different richness. We conclude that, for estimation, the richness of correlation matrices is the key to the theoretical convergences as well as the finite sample performance.

A Comparison Of Batch Versus Iterative Approaches To Vertex Nomination

✦Minh Tang, Johns Hopkins University, Clark Hall 319, 3400 N. Charles St, Baltimore, MD 21218, *mtang10@jhu.edu*; Glen Coppersmith, Johns Hopkins University; Carey Priebe, Johns Hopkins University

Key Words: sequential analysis, attributed graphs

Let G be an attributed graph, i.e., a graph whose vertices and edges have attributes in some discrete sets L_V and L_E, respectively. Suppose that we observe the edge attributes for all the edges and the vertex attributes for a subset of the vertices and that all of these vertices have the same attribute, say 1. The vertex nomination problem is then concerned with nominating a set of vertices whose (unobserved) attribute is most likely to be 1. The nomination of a single vertex can be done using a variety of techniques, one of which is by computing a simple adjacency statistic T(v) for each vertex v with unknown attribute and nominating the vertex v^{*} whose T(v^{*}) is maximum. We investigate the difference between a batch and an iterative approach to vertex nomination that employ these T(v). We aim to show, under a simple model of attributed graphs construction, that depending on the probability that the attribute of a nominated vertex is indeed 1, the batch approach will be better than, comparable to, or worse than the iterative approach.

Hierarchical Characterization Of Loopy Vascular Networks

✦Marcelo Osvaldo Magnasco, Rockefeller University, 1230 York Avenue, Box 212, New York, NY 10065 United States, *mgnsclb@rockefeller.edu*; Eleni Katifori, Rockefeller University

Key Words: Venation, Network, Optimal, Transport

Two dimensional vascular networks, such as leaf veins, retinal vasculature or the surface arteriole network of the cerebral cortex, are characterized by the presence of dense, hierarchically organized loops. Recent work has shown that networks optimized to be resilient to damage, or to spatio-temporal fluctuations in demand display this characteristic recurrent loopy structure. However, detailed morphological comparisons are hindered by the lack of an adequate statistical characterization of such networks. We show that a procedure based on recursive merging of planar facets by breaking the thinnest link in the network results in a well-ordered hierarchy, and that the corresponding tree can be characterized through Horton-Strahler-like laws, permitting detailed morphological characterization of tree species and comparisons between theory and experiment.

Spectral Analysis Of Network Connectivity For Children'S Narrative Comprehension

✦Xiaodong Lin, rutgers university, 252 levin, 94 rockefeller rd., piscataway, NJ 08854 United States, *xiaodonglin@gmail.com*

Key Words: fMRI, Dynamic Bayesian network, Brain Connectivity, Spectral Coherence

In this talk, we present multivariate spectral analysis of fMRI data for a narrative comprehension experiment involving 313 children. In the first part of the study, we apply a Spectral Bayesian Network approach with model averaging to learn the connectivity network underlying active brain regions identified by group ICA. Unlike ordinary Bayesian Networks or Dynamic Bayesian Networks, our method captures the temporal dependency of the entire fMRI time series between brain regions using the spectral density matrices obtained in the frequency domain. A Bayesian model averaging method is applied to select the optimal network structure from a pool of candidates. In the second part, we study the effects of gender and age on the connection strengths between these brain regions. We compute both the pairwise spectral coherence (measuring overall connection strength) and partial spectral coherence (measuring direct link strength) for each subject and then an analysis of covariance is performed. Our method is able to evaluate the impact age and gender have on children's brain development for their narrative comprehension.

A New Statistics For Testing Covariance Struture Of The High Dimension Random Vector

✦Danning Li, University of Minnesota Statistics School, 313 Ford Hall 224 Church St., Minneapolis, MN 55455, *lixx0700@umn.edu*; Tiefeng Jiang, University of Minnesota Statistics School

Key Words: Covariance Structure, Berry_Essen Bound, Chen_Stein Method, Moderate deviation, Random Matrix, Sample Correlation mareix

Testing covariance structure is of significant interest in many areas of statistical analysis. Motivated by these applications, we use another statistic instead of the coherence statistic for testing independence if the p-variates of the population and prove the law of large numbers and the limiting distribution under setting where p can be much larger than n. We also obtain $O(\frac{\log p^{3/2}}{n^{1/2}})$, which is much faster than $O(\frac{1}{\log n})$. We then consider testing the bandedness of the covariance matrix of a high dimensional Gaussian distribution which includes testing for independence as a special case.

137 Modeling of Networks and Information Systems for Defense Applications

Section on Statistics in Defense and National Security, Section on Statistical Computing, Section on Physical and Engineering Sciences, Section on Risk Analysis, Scientific and Public Affairs Advisory Committee

Monday, August 1, 8:30 a.m.–10:20 a.m.

Pattern Discovery And Anomaly Detection In Sensor Networks

✦James Shine, US Army ERDC, , *James.A.Shine@usace.army.mil*; James E Gentle, George Mason University

Key Words: sensors, pattern discovery, anomaly detection

This paper describes an extension of sensor network research performed last year. We have looked in more detail at such issues as differences between groups of sensors, effects of perturbing the data, and modeling normal distributions through the compilation of statistics on sensor activations and interarrival times. We have also modeled different time intervals as independent Poisson processes to look more specifically at times of day that show anomalies, and we have also compared days of the week and times of year. We have also begun work on spatial grouping of sensors and dynamic updating of thresholds and intervals to distinguish anomalies. Results will be presented and discussed.

Variational Methods For Control Over A Directed Graph With Applications To Shipboard Power Management

✦Martin Heller, MAH, 4621 Lemongrass Lane, Durham, NC 27713, *maheller@gmail.com*

Key Words: Stochastic Control, Variational Methods, Energy Efficiency, Reduce Fuel, Optimal Control

Variational methods were used to derive solutions for the stochastic unit commitment problem for directed graphs with consumer, producer and storage nodes. By creating a proper set of basis functions, the calculus of variations solution was shown to be described by a small number of parameters thus reducing the solution space from an infinite dimensional function space down to a finite dimensional Euclidian space. This reduction in complexity makes the variational solution simple to implement and scalable to large systems. The advantages of the variational method is that the solution manifold describes exact local minimum using the exact dynamic functions of the power system and adaptations to real time inline control can be performed us-

ing inexpensive computers. Simulations of the solution are presented which show performance remarkably close to the theoretical minimum attained using an oracle. Further simulations of models for a shipboard power system shows improvements are expected anywhere between 0 and 20%.

A Robust Alternative For Spoof Detection-Using Gmm

✦Umashanger Thayasivam, Rowan University, 36 D, Aspen Hill, Deptford, NJ 08096 USA, *thayasivam@rowan.edu*; Ravi P Ramachandran, Rowan University; Sachin Shetty, Tennessee State University

Key Words: spoof detection, GMM, biometric, robust, SVM, HMM

Biometric technologies have an essential role in assuring and safeguarding personal, national and global security. Such is the value of the assets or information that they protect, biometric systems present a serious and growing target for criminal attack. One form of attack involves so-called 'spoofing' where a person attempts to masquerade as another by falsifying data in order to gain an illegitimate advantage. Alarmingly, as widely acknowledged in the literature, the threat to biometric technologies from spoofing attacks is all too real. A spoof detection system (SDS) is imperative to counter the possibility of hackers using record-and-play spoof attacks to compromise the biometric speaker recognition system. To implement the SDS, we will first evaluate the effectiveness of the popular stochastic based anomaly detection models namely, Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), and Support Vector Machines (SVM). We also propose a robust alternative with L2E distance approach based on GMM. We will perform the experiments on the same testbed and choose the appropriate model based on the following low false positives, quick spoof detection and response time.

Optimal Weighting For A Joint Optimization Of Fidelity And Commensurability Framework For Tests Of Matchedness

✦Sancar Adali, Johns Hopkins University, 3400 North Charles Street, 100 Whitehead Hall, Baltimore, MD 21211 United States, *sadali1@jhu.edu*; Carey Priebe, Johns Hopkins University

Key Words: Multidimensional scaling, manifold matching, disparate sources, fidelity and commensurability, information fusion, Canonical Correlation

For matched data from disparate sources (objects observed under different conditions), optimality of information fusion must be defined with respect to inference task at hand. Defining the task as matched/unmatched hypothesis testing for dissimilarity observations, the forthcoming Manifold Matching paper by Priebe et al. presents an embedding method based on joint optimization of fidelity (preservation of within-condition dissimilarities between observations) and commensurability (preservation of between-condition dissimilarities between observations of an object). We investigate the tradeoff between fidelity and commensurability by varying weights in weighted MDS of omnibus dissimilarity matrix. Optimal (defined with respect to power of test) weights of MDS optimization correspond to an optimal compromise between fidelity and commensurability. The two extremes of this tradeoff are commensurability optimization prioritized over fidelity optimization and vice versa. Results indicate optimal weights are different than equal weights of commensurability and fidelity and our wMDS scheme provides significant improvements in test power compared to embedding via unweighted MDS.

Evaluation Of A Resilience Training Program: The Influence Of Program Mediators On Service Members' Resilience

✦Weimin Zhang, Samueli Institute, 1737 King Street, Suite 600, Alexandra, VA 22314, *wzhang@siib.org*; Salvatore V Libretto, Samueli Institute; Dawn Wallerstedt, Samueli Institute; Joan Walter, Samueli Institute

Key Words: linear and nonlinear regression, path models, mediation model, principal stratification, resilience

Resilience is the capacity to cope with or adapt to significant risk and adversity and to recover quickly from stressful change or misfortune. A skills-based resilience training program designed to aid service members in managing combat stress and enhancing resilience to maximize performance was delivered to over 4,000 soldiers in two Brigade Combat Teams at three time points in 2009 and 2010. In this set of analyses we utilize data obtained from soldiers to explore relevant questions such as: Does the resilience training program have an effect on service members' resilience? Do the constructs targeted by the cognitive and skills training program (i.e., knowledge of perceived control, self-enhancement skills and positive emotions) have an effect on resilience? Do the constructs directly affect resilience or are they mediated by the use of skills? Do ethnicity, army rank, and deployment status work with the constructs of the program to have an impact on resilience? Mediation models will be explored to help explain the differential effects. Hypotheses will be tested using linear and nonlinear regression path models.

Information Systems Success Factors

✦Ayodele Mobolurin, Howard University, 2600 Sixth Street, NW, Washington, DC 20059 USA, *amobolurin@howard.edu*; Mohammad Abul Quasem, Howard University, Dept. of Information Systems & Decision Sciences, School of Business

Key Words: Success Factors, Information Systems, Productivity

This paper reviews the different measures of Information Systems success and provides a classification the for selection of an appropriate measures and the identification of factors that affect Information Systems success. The rate of innovation is very fast and the identification of appropriate measures of success factors is of critical importance to any organization in a very competitive environment

Service-Request Distributions In Business Processes And Their Control-Oriented Applications

Genady Grabarnik, St. John's University; ✦Yefim Haim Michlin, Technion - Israel Institute of Technology, Hashiqma st. 6/19, POB 7626, Nesher, 36812 Israel, *yefim@technion.ac.il*; Laura Shwartz, IBM TJ Watson Research Center

Key Words: service, request process, sequential test, comparison test

In this paper we analyze statistical characteristics of service requests for several different real business processes, with a view to identifying the laws which govern frequently-occurring distributions of the requests. We argue that the high uncertainty in the service processes necessitates a measuring framework with low dependency on variance. These distributions and their characteristics serve as a base for modeling, optimization and control of the processes. The paper describes how to design comparison tests based on Wald's sequential analysis, for estimating the value and evaluating the deviation of the main measurements of a business process: yield, processing time, throughput and error rate. The obtained characteristics are applied in the design of sequential tests for decision support, effectiveness of modifications, and such updates as resource re-allocation, training or retraining of personnel, etc. The paper concludes with examples illustrating the suggested framework in operation for a major service provider in real life situations.

138 Methods for Genetic and Isotopic Environmental Data ■●

Section on Statistics and the Environment, Scientific and Public Affairs Advisory Committee

Monday, August 1, 8:30 a.m.–10:20 p.m.

Stochastic Modeling And Simulation Of Ascertainment Bias In Population Genetics Of Inter-Species Studies

✦Biao Li, Rice University, Department of Bioengineering, Rice University, 6500 Main Street, Suite 135, Houston, TX 77030, *li.biao@rice.edu*; Marek Kimmel, Rice University

Key Words: ascertainment bias, population genetics, inter-population variability, microsatellites, coalescence, forward-time simulation

In population genetics studies, investigators often calculate the inter-population variability at certain homologous loci, e.g. microsatellites and single nucleotide polymorphisms (SNPs), in closely related species. Ascertainment bias arises when a polymorphic locus is selected under a biased discovery process in species 1 and then typed in another species 2. However, allele variability can be influenced, besides the ascertainment bias, by genetic forces such as drift and mutation as well as time-variable demography. Based on the application of stochastic process and Wright-Fisher-Coalescence theory with constant genetic factors, we construct a model to discern ascertainment bias from effects of other forces and aim to implement it to more realistic demographic scenarios with changeable factors. We also use simuPOP, an individual-based forward-time population genetics simulation environment, to test the validity of the model. Simulation results agree very well with the modeling results. By fitting the model to experimental data, we characterize the effect of the ascertainment bias and estimate the impact of demography and genetic forces in inter-species studies.

Sparse Selection Of Multivariate Responses In Association Analysis: Study Of The Effect Of Air Particles On Dna Methylation In A Gene Set

✦Tamar Sofer, Biostatistics Department, HSPH, United States, *tsofer@hsph.harvard.edu*; Arnab Maity, North Carolina State University; Brent Coull, Biostatistics Department, HSPH; Andrea Baccarellli, Environmental Health Department, HSPH; Joel Schwartz, Environmental Health Department, HSPH; Xihong Lin, Harvard School of Public Health

Key Words: Gene and Environment, Epigenetics, Sparsity, Variable Selection, Canonical Correlation Analysis, Principal Component analysis

In environmental epigenetics, one is interested in studying the effects of environmental exposures on DNA methylations in a genetic pathway, which often consists of a large number of genes and many of them are likely to be unaffected by exposures. We develop three Sparse Outcome Selection (SOS) methods for modeling the association between multivariate responses in a genetic pathway and exposures by selecting a subset of genetic outcomes in the pathway whose linear combination yields the highest correlation with a linear combination of exposure variables. We propose SOS Principal Components Analysis, which is a semi-supervised correlation method; SOS Canonical Correlation Analysis (CCA) and the step-forward CCA, which are supervised correlation methods. We investigate three criteria for selecting the tuning parameter, which include prediction correlation, Bayesian Information Criterion, and Correlation Information Criterion, which we developed. We compare the performance of these methods with existing methods via simulations, and apply the methods to the Normative Aging data to study the effects of exposure to airborne particulate matter on DNA methylation in the asthma pathway.

Spatial Models For Bird Origin Assignment Using Genetic And Isotopic Data

✦Colin Witter Rundel, UCLA, 3410 Club Dr, Apt 5, Los Angeles, CA 90064 US, *crundel@ucla.edu*; John Novembre, UCLA; Michael Wunder, University of Colorado Denver; Andrew Schuh, Colorado State University

Key Words: ecology, spatial, bayesian, genetic, isotope, computation

In bird species with large migratory ranges it is often of interest to determine the spatial origin of a particular individual or group of individuals. These data has traditionally been collected through direct observation, e.g. banding or satellite tracking, which tends to be difficult, time consuming, and expensive. Recent work has used genetic or isotopic data to infer these spatial origins with some success, however these results tend to lack sufficient specificity to be useful for many applications. Our work seeks to improve the efficacy of these existing methods by improving spatial predictions of bird origin by combining genetic and isotopic models. This talk will focus on the underlying details of the Bayesian spatial models including specific improvements we have made, as well as approaches for efficient computation. We will also discuss the inclusion of additional spatial information in refining prediction, such as the use of species distribution models as prior information for bird origin.

A Bayesian Framework for Stable Isotope Mixing Models

✦Erik Barry Erhardt, Mind Research Network, 1101 Yale Blvd. NE, Albuquerque, NM 87106, *erik@statacumen.com*; Edward Bedrick, University of New Mexico Health Sciences Center

Key Words: animal ecology, basic mixing model, MCMC, resource utilization

Stable isotope sourcing is used to estimate proportional contributions of sources to a mixture, such as in the analysis of animal diets and plant nutrient use. Statistical methods for inference on the diet proportions using stable isotopes have focused on the linear mixing model. Existing frequentist methods provide inferences when the diet proportion vector can be uniquely solved for in terms of the isotope ratios. Bayesian methods apply for arbitrary numbers of isotopes and diet sources but existing models are somewhat limited as they assume that source means or discrimination are estimated without error or that isotope ratios are uncorrelated. We present a Bayesian model for the estimation of mean diet that accounts for uncertainty in source means and discrimination and allows correlated isotope ratios. This model is easily extended to allow the diet proportion vector to depend on covariates. Two examples are used to illustrate the methodology.

Mix Kin: Delineate Structure In Population Genetics

Arun Sethuraman, Department of Ecology, Evolution and Organizmal Biology, Iowa State University; Karin S Dorman, Iowa State University; Fredric Janzen, Department of Ecology, Evolution, and Organismal Biology, Iowa State University; ✦Wei-Chen Chen, Department of Statistics, Iowa State University, Snedecor Hall, Ames, IA , *snoweye@iastate.edu*

Key Words: Population Genetics, Mixture, Admixture, Model-based Clustering

Inferring diversity of ecosystem, determining number of subpopulation, and classifying species for conservation are important to population genetics. To delineating structure of populations using genetic data, we introduce a model-based clustering approache and impose multinomial distributions to mixture and admixture models. The mixture model assumes individuals were identically selected from K populations, while the admixture model allows individuals were selected with different probabilities. The EM algorithm with analytic solutions is developed for estimating parameters of both models, and implemented in our R package, MixKin. We describe its application in studying the population subdivision of the endangered North American reptile, the Blanding's Turtle (Emys blandingii), sampled across four Midwestern states of Iowa, Illinois, Nebraska and Minnesota. The dataset have 212 turtles genotyped in eight microsatellite loci with missing values. Analysis of population structure using MixKin indicates the most likely number of subpopulations tested by elaborated bootstrap procedures.

Gene Expression Profiling On The Classification Of Androgenic Endocrine Disrupting Chemicals--Quart Medaka Medium

✦Ping-Shi Wu, Lehigh University, 14 E. Packer Avenue, Bethlehem, PA 18015, *psw205@lehigh.edu*

Key Words: Classification, Gene Expression, Endocrine Disrupting Activity

Endocrine disrupting chemicals (EDCs), that are often found in municipal wastewater, compromise development and/or sexual maturation of aquatic organisms by mimicking or antagonizing the actions of natural hormones. A bioassay using model organism can be a powerful tool as it enables us not only to predict presence of EDCs in water through biological responses but also to assess potential impacts on aquatic organisms due to EDCs exposure. Here we demonstrate the possibility of EDC classification based on gene expression profiling by microarray technique using medaka as a sentinel organism. In this study, successful classification of estrogenic effect on a validation set by using both support vector machine and Fisher's linear discriminant analysis, achieving 9% (1/11) misclassification rate. This discovery shed a new light on the screening of environmental water with respect to androgenic, estrogenic and thyroidogenic effects, which can be easily extended to screening/monitoring endocrine disrupting activities.

139 Recent Advanced in the Analysis of Binary Data

Section on Statistics in Epidemiology, Section on Health Policy Statistics

Monday, August 1, 8:30 a.m.–10:20 a.m.

Exact Inference For The Common Odds Ratio In A Series Of 2 By 2 Tables And Its Applications To The Cases Of Rare Events

✦Dungang Liu, Rutgers University, 900 Davidson Rd, Apt 77, Piscataway, NJ 08854, *dungang@stat.rutgers.edu*; Regina Y. Liu, Rutgers University; Minge Xie, Rutgers University

Key Words: meta-analysis, continuity correction, rare events, exact inference, odds ratio, zero total study

This paper proposes a simple meta-analysis method for the common odds ratio in a series of independent 2 by 2 tables based on combining confidence distribution (CD) framework. Almost all commonly used meta-analysis methods rely on large sample approximation. When dealing with rare events, such approximation is nevertheless far from appropriate. In addition, these methods often requires artificial imputing to zero cells, which is known to have an uncertain impact on final conclusions. To circumvent these pitfalls, we propose to combine the exact test result of each study. The combined result is summarized as a function over the parameter space, which is easy to use for overall exact inference. This novel approach enables us to incorporate the studies having zero cells without artificial imputing. Asymptotic efficiency of the proposed approach is established. Numerical studies using simulated and real data show its superiority over existing popular methods.

Type I Error Rate Of Non-Inferiority Trials For A Dichotomous Outcome And Historical Comparator Group

✦Caleb Andrew Bliss, Boston University, 801 Massachusetts Avenue, 3rd Floor, Boston, MA 02118, *cbliss@bu.edu*; Joseph Massaro, Boston University

Key Words: Non-inferiority, Type I Error Rate, Historical Controls

Non-inferiority trials are commonly used to evaluate whether experimental treatments, compared to active controls, yield effects not worse than a pre-specified margin. Usually subjects are randomized to both treatment groups and methods for hypothesis evaluation are well defined. In some cases a historical control of data collected from a previously conducted trial is used for comparison. For a dichotomous outcome a natural hypothesis test is a two-sample test evaluating the null hypothesis of inferiority. While this is an application of the normal

approximation for a binomial outcome under the null hypothesis of a non-zero risk difference, the properties of the hypothesis test have not been examined when applied with a historical comparator group. In this study we simulated data with a dichotomous outcome, a historical comparator, and a single experimental treatment group. We evaluated the type I error rate of the two-sample hypothesis test across a range of effect sizes and margins. The type I error rate was deflated. We examine alternatives to correct the test size.

Estimating Relative Risks For Common Binary Outcomes Using Bias-Corrected Sandwich Estimator

✦Wansu Chen, Kaiser Permanente Southern California, 1026 Panorama Drive, Arcadia, CA 91007 USA, *wansu.chen@kp.org*; Feng Zhang, Kaiser Permanente Souther California; Michael Schatz, Kaiser Permanente Southern California; Zoe Li, Kaiser Permanente Southern California; Robert S Zeiger, Kaiser Permanente Southern California

Key Words: common binary outcome, relative risk, odds ratio, epidemiology, bias-corrected sandwich estimator

When the outcome of a study is binary, the most common method to estimate the effect of an exposure factor is to calculate odds ratio (OR) as an estimate of relative risk (RR) using a logistic regression. However, when the disease prevalence is high (>10%), OR is no longer an acceptable estimate for RR. Several methods have been used to estimate RR directly. These include the robust Poisson model, the COPY method based on log-binomial regression and SAS NLP. We evaluated the performance of a pseudo-likelihood based method with a bias-corrected sandwich estimator (BCSE) that is available in SAS. Simulation was conducted with and without covariates. Compared to the three existing methods, the coverage of 95% confidence intervals (CI) for the BCSE method seemed to be over conservative in small samples, nevertheless they were similar to those of other methods in moderate or large samples. Although comparable among all the methods being evaluated, bias tended to be higher among scenarios with higher RR and higher disease prevalence. Our findings suggest that the BCSE method could be an alternative method in epidemiological or clinical studies in which sample size is moderate or large.

A Goodness-Of-Fit Test Of Logistic Regression Models For Case-Cntrol Data With Measurement Errors

✦Ganggang Xu, Department of Statistics, Texas A&M Unviersity, College station, College station, TX 77843, *gang@stat.tamu.edu*; Suojin Wang, Department of Statistics, Texas A&M Unviersity

Key Words: Case-control study, Conditional score, Empirical likelihood, Logistic regression, Measurement error

We study the problem of goodness-of-fit tests for logistic regression models for case-control data when some covariates are measured with errors.We first study the applicability of traditional test methods for this problem by simply ignoring measurement errors and show that in some scenarios they are still effective despite the inconsistency of the parameter estimators. We then develop a test procedure based on Zhang (2001) that can simultaneously test the validity of using logistic regression and correct the bias in parameter estimators for case-control data with nondifferential classical additive normal measurement error. Instead of using the information matrix considered by Zhang (2001), our test statistic uses a collection of preselected functions to reduce dimensionality. Simulation studies and an application are carried out to illustrate the usefulness of the test.

Alternative Estimators Of Odds Ratio

✦Zahirul Hoque, United Arab Emirates University, Dept of Statistics, Faculty of Business and Econom, PO Box 17555, Al Ain, International UAE, *Zahirul.Hoque@uaeu.ac.ae*; Atanu Biswas, Indian Statistical Institute

Key Words: Odds ratio, estimators, efficiency, preliminary test

This study considers several alternative estimators of Odds Ratio (OR). The estimators considered are the sample OR, Mantel-Haenszel estimator (MHE) and the preliminary test estimator (PTE). The properties of the estimators are studied with respect to bias and mean squared error. It is revealed that the PTE is biased. However, with respect to the mean squared error property, under certain conditions this estimator is more efficient than the other two estimators.

Risk Estimation For Non-Rare Events

✦Shailendra N. Banerjee, Centers for Disease Control and Prevention, 4770 Buford Highway, Atlanta, GA 30341, *snb1@cdc.gov*

Key Words: log-binomial, iterative estimation, converge, non-rare event

A fitted logistic regression model can estimate risks adjusted for confounders, but it is known to over-estimate the risks when the outcome event is common or non-rare with an incidence or prevalence rate of roughly 10% or more. One of the alternatives is the log-binomial model, which is a generalized linear model with log-link and binomial errors. Since, this model estimates relative risk and also because odds ratio does not approximate to relative risk for non-rare events, this model is more appropriate compared to logistic regression in this situation. However, one of the known problems with the use of this model is that the iterative estimation algorithm may fail to converge. We used logistic regression, log-binomial and Poisson regression models on a prostate dataset listed by Collett (1991), a non-rare event of 38% in this dataset. Both log-binomial and Poisson model gave smaller and narrower range of risks compared to that of logistic regression model. The log-binomial regression model also resulted in some non-convergent, out-of-bounds predicted probabilities. This will be further investigated with simulated and more observed data.

Estimation Of A Marginal Causal Odds Ratio In A Matched Case-Control Design

✦Emma Persson, UmeÂ University, UmeÂ, 901 87 Sweden, *emma.persson@stat.umu.se*; Ingeborg Waernbaum, UmeÂ University

Key Words: causal inference, retrospective sampling, confounding

A common practice in the analysis of the effect of an exposure of interest on the development of a rare disease using case-control data is the estimation of odds ratios. A conditional odds ratio can help a clinician to decide whether or not a treatment is beneficial for a particular patient, while a marginal odds ratio can be used to assess the effect of a treatment in the population as a whole. Whereas statistical develop-

ment has to a large extent focused on the former the latter is a parameter more relevant for decision makers. In this paper we compare estimators of the marginal causal odds ratio. Recently, for matched case control designs, targeted maximum likelihood estimators of marginal causal parameters have been proposed. Here, comparisons are made to standard estimators of the odds ratio. Also, an estimator of the marginal causal odds ratio for unmatched case-control designs is proposed. The estimator is based on an intercept adjusted logistic regression model. The finite sample performances of the estimators are highlighted in simulations. The estimators are also applied to data where the effect of socioeconomic variables on the risk of type 1 diabetes is studied.

140 Novel Approach for Infection Disease

Section on Statistics in Epidemiology, ENAR, Scientific and Public Affairs Advisory Committee

Monday, August 1, 8:30 a.m.–10:20 a.m.

Modeling A Within-School Contact Network To Understand Influenza Transmission

✦Gail Elizabeth Potter, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M2-C200, Seattle, WA 98109-102, *gail.potter@gmail.com*

Key Words: social network, epidemic, influenza, agent-based simulation model, contact network

Influenza pandemics pose a serious global health concern. When a new influenza virus emerges with pandemic potential, large-scale simulation models are used to estimate key parameters and compare intervention strategies. Many epidemic models approximate social contact behavior by assuming "random mixing" within mixing groups (e.g., homes, schools, workplaces), and the impact of more realistic social network structure on estimates is an open area of exploration. We develop a realistic model of social contact behavior within a high school using friendship network data and a survey of contact behavior. We perform disease simulations to investigate the impact of dynamics in contact behavior on the epidemic and find that contact behavior is adequately represented with a static contact network. We also simulate a targeted antiviral prophylaxis intervention strategy and a grade closure intervention strategy. We find important differences in the estimates of epidemic outcomes and intervention impact obtained from our model and those obtained by a comparable random mixing scenario. Our findings have implications for policy recommendations based on models assuming random mixing.

Estimates Of Intraclass Correlation Coefficients (Iccs) From A Group-Randomized Hiv Prevention Trial In An Hiv-Positive Population In Sub-Saharan Africa

✦Jun Zhang, ICF Macro, 3 Corporate Square NE, Suite 370, Atlanta, GA 30329, *jzhang5@cdc.gov*; Sherri L Pals, Centers for Disease Control and Prevention; Amy Medley, Centers for Disease Control and Prevention; Catherine Nichols, Centers for Disease Control and Prevention

Key Words: Group-randomized trials, intraclass correlation coefficient, HIV/AIDS

Group-randomized trials (GRTs) are often used to test the efficacy of structural interventions or when individual randomization is not feasible. This design requires an estimate of the expected intraclass correlation coefficient (ICC) to estimate the sample size needed, but few ICC estimates from GRTs in HIV/AIDS have been published. This study will evaluate the effectiveness of an HIV prevention intervention among 3,547 HIV+ patients in HIV Care and Treatment clinics in Kenya, Namibia and Tanzania. We used baseline data from the study to estimate ICCs for 3 dichotomous variables related to unprotected vaginal sex in the past 3 months, taking HIV medications (antiretroviral and/or prophylaxis) and missed HIV medications in the past 30 days. The ICC for unprotected vaginal sex was 0.000005, and ICCs for missed medication and on medications were 0.036 and 0.184, respectively. With 200 patients per clinic, the ICC for missed medications would result in variance of at least 8 times as large as expected with independent observations. Our analyses contribute to knowledge regarding ICCs for outcome variables related to HIV/AIDS interventions and can be useful in designing future trials.

Surgical Site Infection Measure For National Reporting

✦Yi Mu, CDC, 625 Montauk Way, Alpharetta, GA 30022, *hrb3@cdc.gov*; Jonathan Edwards, CDC; Teresa C. Horan, CDC; Scott Fridkin, CDC

Key Words: Quality, hospital, hierarchical, surgical site infection, risk adjusment

The CDC's National Healthcare Safety Network (NHSN), used by CDC and its partners for surveillance of healthcare-associated infections, provides a Standard Infection Ratio (SIR) to participating hospitals to help promote healthcare quality improvement. The SIR is the ratio of the observed surgical site infection (SSI) incidence divided by the expected SSI incidence. To calculate the expected SSI incidence for a hospital, the probability of SSI for each surgical patient are summed; the individual probabilities can be calculated by using procedure-specific risk-adjusted models. Procedure-specific risk-adjusted models were derived using 2006-2008 NHSN data, which contained 62,782 colon surgeries and 54,877 abdominal hysterectomies. Models for procedure-specific risk adjustment were developed using step-wise logistic regression and later validated using bootstrap sampling. The reliability-adjusted SIR was obtained by using hierarchical modeling techniques to shrink the observed SSI incidence toward the expected incidence given hospital procedure volume and risk factors. The resulting adjusted SIRs are more comparable and stable to better measure quality performance.

Estimating The Variation In Malaria Incidence Rate In Heterogeneous Populations

✦Osho O. Ajayi, American University of Nigeria, Yola, Nigeria, *osho.ajayi@aun.edu.ng*

Key Words: Possion mixture distribution, Malaria model, Incidence rate heterogeneity, Malaria data and population diversity

Malaria kills hundreds of thousands of people, including children, around the world every year. While these aggregation of malaria related death figure across age and other important demographic variables make it possible for people actively working to find a plausible solution to the problem to easily comprehend its seriousness, it however hides the fact that in a heterogeneous community, the incidence of reported cases of malaria is highly variable. A good knowledge of the level of this incidence variability among communities can be a useful tool for an effective control policy formulation. This work used a poisson mixture distribution to models the incidence of malaria in a community greatly localized but highly heterogeneous with respect to ethnic background and its influence on daily cultural practice dictating lifestyle. The parameters of the model showed significant variation between groups and may be used a tool for an effective malaria control program deployment in settings with very similar settlement characteristics.

Statistical Approach to Assess the Virulence of Emerging Infectious Diseases

✦Shenghai Zhang, Centre for Communicable Diseases and Infection Control, Ottawa, ON K2J0H3 Canada, *shenghai.zhang@phac-aspc.gc.ca*

Key Words: Case fatality ratio, infectious diseases, Mixed model, nonparametric methods

The case fatality ratio (CFR), the conditional probability of death given infection, has been used to assess the virulence of infection. A time-series method for estimating the CFR of an emerging infectious disease with censored aggregate data is discussed. The approach is based on a mixture model for analyzing data from hospitalization due to outbreak of the disease. The method with data from the 2009 pandemic influenza A(H1N1) is illustrated.

Simulation Studies Of Self-Controlled Case Series Methods In Vaccine Safety Research

✦Guoying Sun, FDA, 20852, *guoying.sun@fda.hhs.gov*; Wei Hua, FDA; Nick Andrews, Health Protection Agency; Caitlin N Dodd, Cincinnati Children's Hospital Medical Center; Silvana A Romio, Erasmus University Medical Center; Hector Izurieta, FDA; Heather J Whitaker, Open University

Key Words: Self-controlled case series, Pseudo-likelihood, post-vaccination, Contraindication

Self-controlled case series (SCCS) method was developed to investigate the association between vaccine and adverse event (AE). When the AE is a contraindication to the vaccine, the SCCS assumption that the event must not alter the exposure process is violated. To investigate this problem, we ran a series of simulations to determine the magnitude of bias and to evaluate different methodologies developed for dealing with this. Three analysis approaches were used to assess the power and accuracy of estimates: 1) the standard SCCS method, 2) post-vaccination follow-up time only with the standard method and 3) the pseudo-likelihood method. The simulations showed that when there was no contraindication to vaccination, the standard method made the best use of all exposure information appropriately and provided higher power; when the contraindication did exist, the pseudo-likelihood method was more appropriate, providing more accurate point estimates. The post-vaccination cases only method worked well when individuals received a single exposure, with a lower power relative to other methods. These results could provide insight into choosing the most appropriate method in real data analysis.

Design Of An Observational Study To Evaluate The Role Of Hiv In Early-Onset Disease In Newborns

✦Elizabeth Zell, CDC, 3166 Bolero Drive, Atlanta, GA 30341, *ezr1@cdc.gov*

Key Words: Propensity Score Methods, Observational Data, Conditional Associations

Frequently in epidemiology, it is important to identify risk factors associated with a specific disease, often from observational studies. Propensity score methods are commonly used in the design of observational studies for causal inference. However, there is also a role for these methods to assess more interesting conditional associations. For example, the question was asked, "Is HIV exposure a risk factor for developing early-onset disease?" We applied propensity score matching to determine if infants born to HIV positive mothers are at increased risk of developing early-onset disease than infants born to HIV negative mothers. For each infant born to an HIV positive mother, we identified an infant born to an HIV negative mother by matching on known risk factors for developing early-onset disease. Propensity score matching allowed us to select an appropriate referent group similar on known or suspected risk factors while blinded to the outcome. The strength of our study was our ability to utilize propensity score methods to better evaluate associations between our comparison groups. Here we illustrate why the matched subset gives the more realistic comparisons between the groups.

141 Analysis of Social Network Data: New Models with Applications to Biology and Health ■●

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

Monday, August 1, 10:30 a.m.–12:20 p.m.

Conditionally Dyadic Independent Models for Longitudinal Analysis of the Effect of Health Traits on Relationships in a Social Network

✦A. James O'Malley, Harvard Medical School, Department of Health Care Policy, 180 Longwood Avenue, Room 301C, Boston, MA 10017 USA, *omalley@hcp.med.harvard.edu*; Sudeshna Paul, Harvard Medical School

Key Words: Dyadic independence, Health, Lagged predictors, Latent variables, Longitudinal model, Social network

This talk considers mixed effect (hierarchical) models for longitudinal analysis of individuals' relationships. We begin by considering some of the troublesome issues that arise in cross-sectional analysis of relational data and use this to motivate longitudinal studies. Particular focus will be on the implications of using observed versus latent variables to account for network dependencies involving the pair of individuals in

each dyad and also on the computational methods for estimation. The models will be applied to a large social network and used to study the relationship between individuals' health behaviors and changes in their friendship and spousal relationships.

A Separable Model for Dynamic Networks

✦Pavel Krivitsky, Carnegie Mellon University, , *pavel@stat.cmu.edu*

Key Words: social networks, Longitudinal, Exponential random graph model, Markov chain Monte Carlo, Maximum likelihood estimation

Models of dynamic networks --- networks that evolve over time --- have manifold applications. We develop a discrete-time generative model for social network evolution that inherits the richness and flexibility of the class of exponential-family random graph models. The model facilitates separable modeling of the tie duration distributions and the structural dynamics of tie formation. We develop likelihood-based inference for the model, and provide computational algorithms for maximum likelihood estimation. We illustrate the interpretability of the model in analyzing a longitudinal network of friendship ties within a school.

Jointly Modeling Homophily in Networks and Recruitment Patterns in Respondent-Driven Sampling of Networks

✦Joseph Blitzstein, Harvard University, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138 USA, *blitzstein@stat.harvard.edu*; Sergiy O. Nesterko, Harvard University Statistics Department

Respondent-driven sampling (RDS) is a network sampling design which is widely and increasingly being used to study properties of individuals in a social network, e.g., HIV prevalance. Understanding and evaluating the estimators obtained through RDS depends heavily on understanding the process of who recruits whom, which in turn depends on homophily in the network. We propose a joint model for the underlying homophily and the RDS recruitment process, and demonstrate how it can be used in RDS estimation, especially in obtaining standard errors from a model-based perspective.

Neighborhood and Network Effects on Health

✦Felix Elwert, University of Wisconsin-Madison, , *elwert@wisc.edu*

Research on the context dependence of health outcomes distinguishes between two different types of social context: patients' neighborhoods of residence and patients' social networks. This paper investigates conceptual difficulties and estimation challenges of separating network effects from neighborhood effects in order to adjudicate their relative contribution to health outcomes. An analysis of neighborhood and network effects in the Framingham Heart Study illustrate the argument.

142 Bayesian model assessment

International Society of Bayesian Analysis, International Indian Statistical Association, Section on Bayesian Statistical Science, Section on Statistics and the Environment

Monday, August 1, 10:30 a.m.–12:20 p.m.

Model Checking in an Expanded Graphical Modeling Framework

✦Andrew Gelman, Department of Statistics, Columbia University, New York, New York, NY 10027 USA, *gelman@stat.columbia.edu*

Key Words: Bayesian data analysis, debugging, posterior predictive check

We discuss how to expand the language of directed acyclic graphs to include model checking, model comparison, and debugging, along with the traditional graphical model task of inference within a model. We are not particularly interested in discrete model choice, model averaging, Bayes factors, etc.; rather, we are working in applied contexts in which we are continually building more complicated models which we need to understand in light of earlier, simpler efforts.

Bayesian Selection of Hidden Markov Random Fields

✦Jean-Michel Marin, Universite Montpellier 2, France, , *Jean-Michel.Marin@univ-montp2.fr*; Lionel Cucala, Universite Montpellier 2, France

Key Words: mixture models, spatial dependences, Potts models, model choice, Chib's method, Laplace approximation

We introduce two techniques in order to select the number of components of a mixture model with spatial dependences. Typically, we consider an image field where the grey-scale values are Gaussian random variables depending on the component of the associated pixel and the components are distributed according to a Potts model. The first method comes from an approximation of the evidence based on the Chib's method. The second one is deduced from an approximation of the integrated completed likelihood using Laplace approximations. We compare these two techniques on real and simulated datasets.

Tba

✦Feng Liang, University of Illinois at Urbana-Champaign, 2417 Stricker Ln, Urbana, IL 61820 USA, *liangf@illinois.edu*

TBA

143 Statistical Modeling at the Internet Scale: Understanding User and Advertiser Behavior ■●

Section on Statistical Learning and Data Mining

Monday, August 1, 10:30 a.m.–12:20 p.m.

Analysis of Advertiser Behavior Using Hidden Markov Models

✦Sangho Yoon, Google Inc, 1600 Amphitheatre Parkway, Mountain View, CA 94043 USA, *shyoon@google.com*; Steve Scott, Google, Inc

Key Words: Hidden Markov Models, MCMC, Customer retention

In Google AdWords, advertisers participate in an auction to show their advertisements to users who come to Google and search for information. Google tries to serve both users and advertisers by providing relevant and high-quality information to users and advertising opportunities to advertisers. In this work, we analyze the path of several advertisers through a multi-dimensional space of success metrics, with the goal of separating advertisers on positive and negative trajectories. The modeling is implemented using hidden Markov models that use latent time-dependent states to capture variation in the time series of multivariate observations.

Large-Scale Social Network Analysis at Google

✦Rachel Schutt, Google Inc, 10025 USA, *schutt@google.com*; Daryl Pregibon, Google Inc.

Key Words: social networks, large-scale, google, algorithms

The study of networks is cross-disciplinary and research is ongoing in fields as diverse as sociology, computer science, mathematics, statistics, physics, epidemiology and biology. We define a network to be a structure of nodes that are connected by edges. We can think of the nodes as representing people or businesses and edges as representing relationships. Data of this type can be found in abundance at Google, and exists on a massive scale.There are a number of very interesting open problems in the field of social network research including but not limited to: modeling social networks, the development of dynamic network models; time series models for networks; prediction-type problems; sampling from networks, and samples of networks; and epidemics or processes on networks. These problems are interesting unto themselves, and the large-scale data aspect adds an additional level of complexity. We will show how we have used the Google-developed large-scale graph computation algorithm and infrastructure, Pregel, to begin to attack some of these problems.

Facebook's Entities Graph

✦Eric Sun, Facebook, Inc., , *esun@facebook.com*

Key Words: Facebook, entities, social networks, crowdsourcing, large-scale

Facebook Community Pages are an attempt to create a catalog of all known entities in the world. These concepts can be added to users' profiles on Facebook, allowing them to express their passions and share their interests with others. In this talk, we discuss the challenges and progress building and maintaining Facebook's social graph of entities--that is, the concepts, places, and things to which our users connect. With hundreds of millions of Pages, problems like deduplication and disambiguation quickly become computationally difficult. We propose several solutions for these problems that can be applied at Facebook's scale, including cleaning up the graph via statistical algorithms, integrating user feedback via crowdsourcing, and inferring entities from text.

Demographic Diversity on the Web

✦Jake M Hofman, Yahoo! Research, 111 W 40th St, 17th Floor, New York, NY 10018 USA, *hofman@yahoo-inc.com*

Key Words: web, browsing, demographic, diversity

To what extent do the online experiences of, for example, men and women, or Whites and African-Americans differ? We address such questions by pairing web browsing histories for 265,000 anonymized individuals with user-level demographic data---including age, sex, race, education, and income. In one of the most comprehensive analyses of Internet usage patterns to date, we make three broad observations. First, while the majority of popular sites have diverse audiences, there are nonetheless, prominent sites with highly homogeneous user bases. Second, although most users spend a significant fraction of their time on email, search, and social networking sites, there are still large group-level differences in how that time is distributed. Finally, the between-group statistical differences enable reliable inference of an individual's demographic attributes from browsing activity. We thus conclude that while the Internet as seen by different demographic groups is in some regards quite similar, sufficient differences persist so as to facilitate group identification.

144 Social Media in Marketing ■●

Section on Statistics and Marketing, Section on Statistical Computing

Monday, August 1, 10:30 a.m.–12:20 p.m.

Online Product Opinions: Incidence, Evaluation, and Evolution

✦Wendy Moe, University of Maryland, 3469 Van Munching Hall, College Park, MD 20742 U.S.A., *wmoe@rhsmith.umd.edu*; David A. Schweidel, University of Wisconsin-Madison

Key Words: Social Media, Online Opinion, Social Dynamics, Ordered Probit, Product Ratings

In this research, we empirically model the individual's decision to provide a product rating and investigate factors that influence this decision. Specifically, we consider how previously posted opinions in a ratings environment may affect a subsequent individual's posting behavior, both in terms of whether to contribute (incidence) and what to contribute (evaluation), and identify selection effects that influence the incidence decision and adjustment effects that influence the evaluation decision. Our results indicate that individuals vary in their underlying behavior and their reactions to the product ratings previously posted. We also show that posted product opinions can be affected substantially by the composition of the underlying customer base and find that products with polarized customer bases may receive product ratings that evolve in a similar fashion to those with primarily negative customers as a result of the dynamics exhibited by a core group of active customers.

Listening in on Online Conversations: Measuring Consumer Sentiment with Social Media

✦David A. Schweidel, University of Wisconsin-Madison, 975 University Ave, 4191B Grainger Hall, Madison, WI , *dschweidel@bus.wisc.edu*; Wendy Moe, University of Maryland

Key Words: Social Media, Online Opinion, Marketing Research

With the explosion of data available online, businesses are increasingly turning to social media as a listening tool. Some firms choose to engage with their customers directly using social media, while others leverage

the insights they glean in other marketing activities. Recent research has examined movement in the volume of social media and the sentiment expressed. Much work, however, focuses on comments contributed to a single online venue. In this research, we probe the differences in sentiment that manifest across multiple types of venues, including blogs, forums and social networks. We also investigate differences in sentiment that may exist between contributors with and without direct experience. Our analysis reveals variation expressed sentiment across venues and based on contributors' experiences, important cautions for those gauging consumer sentiment from social media.

Repeat Purchasing in a Social Network Setting

✦Peter Fader, University of Pennsylvania, 771 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104, *faderp@wharton.upenn.edu*; Eric Schwartz, University of Pennsylvania; Renana Peres, Hebrew University

Key Words: social network, agent-based modeling, stochastic models, repeat purchasing, customer attrition, diffusion models

We study this link between repeat purchasing, customer attrition, and social network structure/activity. Using an agent-based model, we simulate the initial diffusion, ongoing repeat usage, and ultimate customer attrition for a new product in a social network setting. The adoption, repeat purchasing and attrition decisions of individuals in the network depend on the communication dynamics within the social network, and are described using a well-established "buy till you die" probability model that has been augmented to account for network characteristics. To obtain the parameters used in the agent-based model (and to establish the overall validity of our proposed approach), we first estimate this type of model using actual data from a new application on Facebook. We then expand upon this empirical model using agent-based techniques, in order to explore how the overall long term profitability of the network depends on the repeat-purchasing/attrition parameters, and the sensitivity of repeat purchasing/attrition to the network characteristics. We also demonstrate the biases created when using a plain repeat-purchase model that ignores the social network characteristics.

Inter-Media Reactivity: A Conceptual Framework and Methodology for Analyzing Dynamics of New Media

✦Amit M Joshi, University of Central Florida, Orlando, FL 32816 US, *ajoshi@bus.ucf.edu*

Key Words: Advertising, VAR Models, User-Generated Content, Social Media, Motion Picture Industry

The rapid spread of social media has not only brought new forms of media into marketing communication mix, but also meant that messages are commonly transferred between media leading to multiple media dissemination by empowered consumers. Thus, a firm's control over its communication mix has been diminished and traditional managerial and academic knowledge on how to manage the communication mix may soon become obsolete. In the context of these developments, this research introduces the novel concept that all entities in the information space can be characterized by a set of attributes, which we refer to as their Inter Media Reactivity (IMR), that determine how various media react to information regarding the entity and interact when carrying this information. Inter Media Reactivity varies across entities, the cross-media interactions can be asymmetric and the duration of media response can vary across media. IMR can thus help firms determine how different media are likely to react to information about their products that originates in a particular medium, and for how long the other media will remain stimulated.

145 The Interface Between Statistics and IT ■●

Section on Statistical Consulting, Section on Quality and Productivity

Monday, August 1, 10:30 a.m.–12:20 p.m.

Sourcing the IT Infrastructure for World-Class Analytics

✦Maneesh Aggarwal, Travelers Insurance, One Tower Square, PB05-A, Hartford, CT 06183, *maggarwa@travelers.com*

Key Words: Infrastructure, Sourcing, Computing

IT infrastructure is a key enabler in the lifecycle of developing and implementing statistical insights. However, lack of proper computing infrastructure is commonly cited as a key reason for not being to able to maximize the value of investments in the analytic process as well as a contributing cause of statistician dissatisfaction and poor morale. The relationship between the technical disciplines of IT and Statistics is complex. Effective partnership between IT and Statistician teams is important to building the right infrastructure. This paper discusses the nature of the problem and presents an example of deep collaboration between these teams in sourcing a solution that addresses some of the issues.

The Value of a Close Statistics/IT Relationship: A Case Study at Eli Lilly

✦Todd Sanger, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, *sanger@lilly.com*

Key Words: IT, relationship, SDD

Within the pharmaceutical industry, the Statistics and IT organizations are tightly entwined. The Statisticians are very computer savvy and require special programs and IT systems to support their needs. With the explosion of data and subsequent high power computing needs around genomics and claims databases, many statisticians are forced to consider parallel computing and cloud computing to improve performance. Also, given the strict regulatory compliance demands, special computing environments are needed to maintain tight control and access to clinical data, programs, and results with appropriate traceability and electronic signatures. Given these special needs, it is important to have a close working relationship with IT. I will describe the history of the IT/Stat relationship at Lilly. I will particularly discuss how the two organizations worked successfully together to implement SAS Drug Development (SDD) at Lilly. Because of the success of this relationship, Lilly subsequently won a 2008 Bio-IT Best Practice Award for Clinical Research and a 2008 SAS Enterprise Intelligence Award for this implementation of SDD.

The Value of a Close Statistics/IT Relationship: A Scalable Framework

✦Anuwat Raviwongse, Manpower, 100 Manpower Place, Milwaukee, WI 53212, *anuwat.raviwongse@na.manpower.com*

The statistic consultant is often perceived to be tied to one-off projects that answer very specific question at a point in time. This view however marginalizes the ongoing value a statistician or an analytics team can bring to the organization. Business leaders are slowly realizing this disconnect and will expect statistics capabilities to be readily and regularly available for mass consumption. A scalable and workable Information Management Framework can help address the challenge of making statistics ubiquitous. We will first walk through the "typical" information delivery framework then how IT connects to the statistical consultant and the business. This portion will cover governance, the typical information management application structure, skills sets, challenges IT faces on a day to day basis, and some trends in IT. The next portion will be a brief discussion of the

146 Statistical Methods for Risk Prediction Using High Throughput Genomic Data ■

Biometrics Section, ENAR, Section on Statistics in Epidemiology, Section on Risk Analysis

Monday, August 1, 10:30 a.m.–12:20 p.m.

Predicting the Future of Genetic Risk Prediction

✦Nilanjan Chatterjee, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, USA, 6120 Executive Blvd,, Rockville, MD 20852 USA, *chattern@mail.nih.gov*; JuHyun Park, National Cancer Institute; Mitchell Gail, National Cancer Institute

Key Words: genome-wide association study, gene-environment interaction, ROC curve, Complex traits

Although recent genome-wide association studies have led to the identification of many susceptibility loci for a variety of complex traits, the utility of these discoveries for predicting individualized risk has been modest. This talk will examine the potential utility of future risk models that may include additional susceptibility loci as well as non-genetic risk factors. In particular, we will describe methods for estimating number of underlying susceptibility loci for a trait and the distribution of their effect-sizes using data from recent genome-wide association studies. We will then show how such estimates together with existing risk factors for diseases can be used to assess the limits of performance of future prediction models Conversely, we also evaluate the magnitude of effect-sizes and number of risk factors needed for risk models to have substantial discriminatory power and hence have major impact for public health applications.

Risk Prediction with Genome-Wide Association Studies

✦Tianxi Cai, Harvard University, Building I Room 411 / HSPH, 651 Huntington Avenue, Boston, MA 02115, *tcai@hsph.harvard.edu*; Jessica Minnier, Harvard University

Key Words: risk prediction, genetic pathways, high dimensional data

The complexity of the genetic architecture of human health and disease makes it difficult to identify genomic markers associated with disease risk or to construct accurate genetic risk prediction models. Accurate risk assessment is further complicated by the availability of a large number of markers that may be predominately unrelated to the outcome. Standard marginal association based analysis has limited power in identifying markers truly associated with disease, resulting in a large number of false positives and negatives. Simple additive modeling does not perform well when the underlying effects are highly interactive or non-linear. Additionally, these methods do not use information that may be available regarding genetic pathways or gene structure. We propose a multi-stage method relating markers to the risk of disease by first forming multiple gene-sets based on certain biological criteria and then aggregating information across all gene-sets. Prediction accuracy is assessed with bias-corrected ROC curves and AUC statistics. Numerical studies suggest that the model performs well in the presence of non-informative regions and both linear and non-linear effects.

Statistical Issues in Risk Modeling of Complex Diseases Using High-Throughput Data

✦Heping Zhang, Yale University, 60 College Street, New Haven, CT , *heping.zhang@yale.edu*

Genes and environmental factors are believed to underlie the etiology of complex diseases, and to study the risk of complex diseases, high throughput data such as common and rare genetic variants have been generated. Before risk modeling, it is necessary to develop and assess the efficiency of statistical methods and models. Although many real data sets have been collected, they are not ideal for this developmental and evaluation purpose because the true answer is unknown. Thus, simulation data must be generated, must resemble the real data, and must be computationally feasible. In this talk, I will first discuss strategies to generate such data, and then present statistical methods that can be used to identify common and/or rare variants for complex diseases.

147 Statistical Challenges Arising from Next-Generation Sequencing Data ■●

IMS, Biometrics Section, ENAR, International Chinese Statistical Association

Monday, August 1, 10:30 a.m.–12:20 p.m.

New Applications of Next-Generation Sequencing

✦Wing Hung Wong, Stanford University, , *whwong@stanford.edu*

Key Words: sequencing, phasing, hyplotype

The advent of next-generation technology for DNA-sequencing has opened up several important applications in biology and medicine, including ChIP-seq, RNA-seq and the mapping of rare diseases-associated alleles. We believe this is only the beginning and there will be many more novel uses of this technology. Statistical analysis will be central to the development of these new applications. We propose one such new application, namely, to generate completely phased genome sequences in a high throughput manner.

Do Multi-Reads Matter in ChIP-Seq Data Analysis?

✦Sunduz Keles, University of Wisconsin, Madison, 1300 University Avenue, 1245 MSC, Department of Statistics, Madison, WI, WI 53705 USA, *keles@stat.wisc.edu*

Key Words: High throughput sequencing, ChIP-Seq, Mixture models, Mappability, Transcription factors

The introduction of next generation sequencing enabled a myriad of creative ways to answer genome-wide questions. In particular, chromatin immunoprecipitation coupled with sequencing (ChIP-Seq) has become a powerful technique for large scale profiling of transcription factor binding and chromatin modifications. A ChIP-Seq experiment generates millions of short reads. The first step of data analysis is to map reads to reference genome and retain reads that map to a single location in the genome (uni-reads). Restraining the analysis to unireads leads to reduced sequencing depth and further poses a significant challenge for identifying binding locations that reside in regions of genome that have been duplicated over evolutionary time. We investigate the effects of discarding multi-reads in ChIP-Seq data analysis and illustrate that their incorporation can increase sequencing depth up to 20%. We develop a model-based method for taking into account mapping uncertainty in ChIP-Seq data analysis. We quantify and characterize gains from multi-reads in case studies from the ENCODE project, and support our conclusions with both computational and experimental validations.

Statistical Methods for the Analysis of Ribosome Profiling Data

✦Adam B Olshen, UCSF, , *olshena@biostat.ucsf.edu*; Barry Taylor, Memorial Sloan-Kettering Cancer Center; Richard Olshen, Stanford University

Key Words: ribosome profiling, sequencing

During translation messenger RNA produced by transcription is decoded by robosomes to produce specific polypeptides. Recently, a second generation sequencing methodology was developed that measures the position and counts of ribosomes. When combined with corresponding mRNA sequencing data, ribosomal data can give insights into translational efficiency. I will discuss issues that arise when analyzing such data, including normalization and testing hypotheses. I will demonstrate competing methods of analysis on a set of data that are available to the public and by simulation.

Statistical and Computational Methods for the Analysis of Pooled, Targeted, Second-Generation Resequencing Data

✦Hector Corrada Bravo, University of Maryland, College Park, Center for Bioinformatics and Computational Biolog, Biomolecular Sciences Building #296, College Park, MD 20742, *hcorrada@umiacs.umd.edu*

The ability to effectively and cheaply genotype a large number of samples by pooled targeted re-sequencing using second-generation technologies holds exciting promise for clinical and biological applications. In this talk we will report findings on systematic biases in this technology that must be addressed in order to properly carry out this task. First, we report on batch effects in samples where targeted regions are captured by hybridization and discuss the impact of this on SNP and CNV finding. Second, we discuss technical biases in base-calling and their effect on genotyping and allele rate estimation. We will report statistical and computational solutions for this problems ranging from base-calling and quality assurance to allele rate estimation and SNP calling.

148 Medallion Lectures: A journey to ultrahigh dimensional space ●

IMS, International Chinese Statistical Association, International Indian Statistical Association

Monday, August 1, 10:30 a.m.–12:20 p.m.

Medallion Lectures: A Journey to Ultrahigh Dimensional Space

✦Jianqing Fan, Princeton University, Department of Operations Research & Finiancial Eng, Princeton University, Princeton, NJ 08544 USA, *jqfan@princeton.edu*

Key Words: Variable selection, high-dimensional data, penalized method, ISIS

Ultrahigh-dimensionality characterizes many contemporary statistical problems from genomics and engineering to finance and economics. To visit such a vast dimension, large-scale and moderate-scale vehicles are needed for such a venture. We outline a unified framework to ultrahigh dimensional variable selection problems: Iterative applications of vast-scale screening followed by moderate-scale variable selection, called ISIS. The framework is widely applicable to many statistical contexts: from multiple regression, generalized linear models, survival analysis to machine learning and compress sensing. The fundamental building blocks are marginal variable screening and penalized likelihood methods. How high dimensionality can such methods handle? How large can false positive and negative be with marginal screening methods? What is the role of penalty functions? This talk will provide some fundamental insights into these problems. The focus will be on the sure screening property, false selection size, the model selection consistency and oracle properties. The advantages of using folded-concave over convex penalty will be clearly demonstrated. The methods will be illustrated.

149 Advancing Analysis of Event History Data ■●

SSC

Monday, August 1, 10:30 a.m.–12:20 p.m.

Regression Analysis of Longitudinal Data with Dependent Observation Process and Application to Medical Cost Data

✦(Tony) Jianguo Sun, University of Missouri, Columbia, 134E Middlebush Hall, Columbia, MO 65211-6100, *SunJ@Missouri.edu*; Liang Zhu, St. Jude Children's Research Hospital

Key Words: Counting processes, Latent variable model, Longitudinal data analysis

Longitudinal data analysis is one of the most discussed and applied areas in statistics and a great deal of literature has been developed for it. However, most of the existing literature focus on the situation where observation times are fixed or can be treated as fixed constants. This paper considers the situation where these observation times may be random variables and more importantly, they may be related to the underlying longitudinal variable or process of interest. For the problem, we present a joint modeling approach and an estimating equation-based procedure is developed for estimation of possibly time-varying regression parameters. The methodology is applied to a set of medical cost data from an acute myeloid leukemia trial.

Analysis of Reaction Times as Discretely Censored Data

✦Willard John Braun, University of Western Ontario, 1151 Richmond Street, London, ON N6A 5B7 Canada, *braun@stats.uwo.ca*

Key Words: reaction time, discrete censoring, point process, kernel density estimation, intensity function estimation

In visual psychophysics, researchers study the brain mechanisms underlying vision by presenting visual stimuli and obtaining behavioural responses from an observer. In a simple reaction time (RT) experiment, a stimulus is presented, and the time taken for the observer to hit a button is measured. RT indicates the complexity of the operations taking place in the brain, and it is simple to analyze. The simple RT experiment consists of a warning, a uniformly distributed delay, then the briefly flashed stimulus. Therefore, the observer can anticipate the stimulus. A more realistic approach is to present the flashes according to a Poisson process and record the time of each button press. Running the experiment in this way causes new difficulties. For example, if the flashes are too close together in time, it is difficult to tell which flash is associated with which button press, leading to a special form of censoring in which the actual reaction time observations belong to discrete sets. We compare two methods for nonparametrically estimating the reaction time density function: Brillinger cross-intensity function estimates; and iterated conditional expectations.

Simultaneous Inference for Longitudinal Data with Covariate Measurement Error and Missing Responses

✦Wei Liu, York University, 4700 Keele Street, Toronto, ON M3J 1P3 Canada, *liuwei@mathstat.yorku.ca*

Key Words: Bias analysis, Longitudinal data, Measurement error, Missing data, Monte Carlo EM algorithm, Random effects models

Longitudinal data arise frequently in medical studies and it is common practice to analyze such data with generalized linear mixed models. Such models enable us to account for various types of heterogeneity, including between and within subjects ones. Inferential procedures complicate dramatically when missing observations or measurement error arise. In the literature there has been considerable interest in accommodating either incompleteness or covariate measurement error under random effects models. However, there is relatively little work concerning both features simultaneously. There is a need to fill up this gap as longitudinal data do often have both characteristics. In this paper our objectives are to study simultaneous impact of missingness and covariate measurement error on inferential procedures and to develop a valid method that both computationally feasible and theoretically valid. Simulation studies are conducted to assess the performance of the proposed method, and a real example is analyzed with the proposed method.

Joint Modeling of Longitudinal and Survival Data to Estimate Treatment Effects

✦Jeremy Michael George Taylor, University of Michigan, Dept of Biostatistics, 1420 Washington Heights, Ann Arbor, MI 48109, *jmgt@umich.edu*

Key Words: prostate cancer, causal effects, treatment by indication, joint models

Estimating a treatment effect of an intervention from longitudinal observational data, when the treatment is given by indication, is challenging. We utilize a longitudinal and hazard modeling approach to estimate the effect of salvage hormone therapy for patients who are being followed after treatment for prostate cancer. The longitudinal data are prostate specific antigen (PSA) counts and the survival data are time of recurrence. Patients with higher PSA values are at higher risk for recurrence and they also tend to be given hormone therapy. We are interested in estimating the reduction in the hazard of the recurrence of the prostate cancer for receiving hormone therapy for that patient conditional on their longitudinal PSA data compared to not receiving hormone therapy. The joint longitudinal-survival modeling approach requires a model for what PSA would have been had the person not taken hormone therapy. The method will be compared with a sequential stratification propensity score method and marginal structural models.

150 Pragmatic Risk Studies ■●

Section on Risk Analysis, Section on Health Policy Statistics, Section on Statistics in Defense and National Security

Monday, August 1, 10:30 a.m.–12:20 p.m.

Hard-to-Study Risk/Benefit Forecasting Issues

✦Michael E. Tarter, Univeristy of California, 779 University Hall, CA 94720, *tarter@berkeley.edu*

Key Words: experimental subject replacement, loglogistic, lognormal, order statistic, stopping rule, threshold parameter

High on any list of practical difficulties is the need to choose truncated models of asymmetric underlying densities. Even in the univariate case these models must be formulated in terms of four or more parameters; hence the motivation for a new model-free procedure for detecting density edge removal. To detect stump presence, two sets of sample trigonometric moments are compared. The first set contains members of moderate order while the second set contains members of high order. The rationale for this new approach stems from the relationship between the rate of convergence of a curve's Fourier expansion and rate at which the curve's border derivatives approach a common value. For example, for representational purposes when a Normal density with intact tails is expanded a handful of trigonometric moments will

suffice. Conversely, a Venus de Milo-like and/or lopsided bell shaped curve will tend to have a sequence of Fourier coefficients that converges slowly.

Adaptive Reuse of National Public Health Survey Samples for Estimating Hard-to-Locate Population Characteristics

✦Myron J Katzoff, CDC/National Center for Health Sataistics, 3311 Toledo Road, Hyattsville, MD 20782, *mjk5@cdc.gov*

Key Words: adaptive sampling, network sampling, indirect sampling

Adaptive re-use of samples from public health surveys can provide information in emergency situations where: (1) there is an urgent need for information; (2) a sampling frame for a small, but possibly growing, target population is not available but one for a conventional design (i.e., one employing traditional nonparametric sampling methods and design-based estimation procedures) is available; and (3) certain provisions can be made for modifications in data-collection practices. In the early stages of a public health emergency, the target populations are typically small and, therefore, hard-to-locate; adaptive sampling are among the approaches developed for estimating the characteristics of these populations. We examine how adaptive methods may be applied to inform the decision-making that accompanies the tracking and monitoring of interventions.

Adversarial Risk Analysis

✦David Banks, Duke University, Department of Statistical Sciences, DSS Box 90251, Durham, NC 27708, *banks@stat.duke.edu*

Key Words: game theory, Bayes, decision analysis

Classical risk analysis assumes that threats are not intelligently directed; that kind of problem is traditionally treated by game theory. But game theory is an unrealistic guide to human behavior---there is much empirical evidence that people do not usually employ minimaxity. This talk describes a Bayesian approach in which the analyst "mirrors" the decision-making processes of the opponent, taking account of two types of uncertainty: uncertainty in the choice of the opponent, and uncertainty in the outcome conditional on the choice of the opponent. The approach is illustrated in the analysis of an auction and the game La Relance, invented by Borel.

Design of Cost-Effective Experiments

✦ Alexandra Kapatou, Performance Management Associates, 3429 Tanterra Circle, Brookeville, MD 20833 USA, *akapatou@comcast.net*; David Banks, Duke University

Key Words: Factorial Design, Information, Optimality, Response Surface

Conventional experimental design theory ignores the fact that different observations have different costs. When some observations are much cheaper to make than others, then experimenters should seek the design which provides the most information at an affordable price. Such designs are typically unbalanced, but can be easily analyzed by modern software. This paper describes the issues that arise in cost-effective design and response surface analysis, and points out how the results differ from those obtained under traditional alphabetic-optimality criteria. The design of cost-effective experiments is illustrated with examples.

151 Recent advances in statistical machine learning and model selection ■●

Section on Nonparametric Statistics, ENAR, International Chinese Statistical Association, International Indian Statistical Association

Monday, August 1, 10:30 a.m.–12:20 p.m.

Sparse Quantile Regression Approach for Analyzing Heterogeneity in Ultrahigh-Dimension

✦Runze Li, Penn State University, Department of Statistics, 326 Thomas Building, University Park, PA 16802-2111, *ril4@psu.edu*; Lan Wang, University of Minnesota; Yichao Wu, North Carolina State University

Key Words: Quantile Regression, SCAD, Ultrahigh-dimensional data

Ultrahigh-dimensional data is often heterogeneous due to either heteroscedastic variance or other forms of non-location-shift covariate effects. Quantile regression is particularly useful for analyzing data from heterogeneous population. Usually in practice, only a few covariates influence the conditional distribution of the response variable given all candidate covariates. We propose to systematically study sparse quantile regression for ultrahigh-dimensional data. For both computation and theoretic development, it is challenging to deal with both the nonsmooth loss function and the nonconvex penalty function in ultrahigh-dimensional parameter space. We develop a new algorithm to deal with computational issue and theoretically analyze the proposed algorithm. The new algorithm enables us to establish a new formulation of the oracle property for ultrahigh-dimensional data. We further study the sampling properties of the penalized quantile regression for ultrahigh-dimensional data under some regularity conditions which are weaker and more reasonable conditions than the existing ones in the literature.

DD-Classifier: Nonparametric Classification Procedure Based on DD-Plot

✦Jun Li, University of California, Riverside, Department of Statistics, University of California, Riverside, Riverside, CA 92508, *jun.li@ucr.edu*; Juan Antonio Cuesta-Albertos, University of Cantabria, Spain; Regina Y. Liu, Rutgers University

Key Words: Classification, data depth, DD-plot, DD-classifier, nonparametric, robustness

Using the DD-plot (depth-versus-depth plot), we introduce a new nonparametric classification algorithm and call it a DD-classifier. The algorithm is completely nonparametric, and requires no prior knowledge of the underlying distributions or of the form of the separating curve. Thus it can be applied to a wide range of classification problems. The algorithm is completely data driven and its classification outcome can be easily visualized on a two-dimensional plot regardless of the dimension of the data. Moreover, it is easy to implement since it bypasses the task

of estimating underlying parameters such as means and scales, which is often required by the existing classification procedures. We study the asymptotic properties of the proposed DD-classifier and its misclassification rate. Specifically, we show that it is asymptotically equivalent to the Bayes rule under suitable conditions. The performance of the classifier is also examined by using simulated and real data sets. Overall, the proposed classifier performs well across a broad range of settings, and compares favorably with existing classifiers. Finally, it can also be robust against outliers or contamination.

From Statistical Learning to Game-Theoretic Learning

✦Alexander Rakhlin, University of Pennsylvania, 3730 Walnut St, Philadelphia, PA 19104 US, *rakhlin@gmail.com*

Statistical Learning Theory studies the problem of estimating (learning) an unknown function given a class of hypotheses and an i.i.d. sample of data. Classical results show that combinatorial parameters (such as Vapnik-Chervonenkis and scale-sensitive dimensions) and complexity measures (such as covering numbers, Rademacher averages) govern learnability and rates of convergence. Further, it is known that learnability is closely related to the uniform Law of Large Numbers for function classes. In contrast to the i.i.d. case, in the online learning framework the learner is faced with a sequence of data appearing at discrete time intervals, where the data is chosen by the adversary. Unlike statistical learning, where the focus has been on complexity measures, the online learning research has been predominantly algorithm-based. That is, an algorithm with a non-trivial guarantee provides a certificate of learnability. We develop tools for analyzing learnability in the game-theoretic setting of online learning without necessarily providing a computationally feasible algorithm. ** This is joint work with Karthik Sridharan and Ambuj Tewari. **

Various Aspects of Model Selection in Functional/Longitudinal Data Analysis

✦Naisyin Wang , University of Michigan, University of Michigan, 447 West Hall 1085 S. University,, Ann Arbor, MI 48109-1107, *nwangaa@umich.edu*

Key Words: functional data analysis, longituidnal data analysis, basis function

In functional and longitudinal data analysis, there are various factors that might determine the quality of the outcomes. In this talk, we will discuss similarities as well as differences between handling functional and longitudinal data. Issues such as the choices of basis functions, model selection and accommodating heterogeneity features in the data will be addressed. We will illustrate our findings through analysis of simulated and field data sets. Part of the work is joinly done with RJ Carroll, Y. Li and JH Zhou.

152 A Summary of the 2010 Census Program of Evaluations and Experiments with an Eye Toward the 2020 Census ■

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

Monday, August 1, 10:30 a.m.–12:20 p.m.

Summary of the 2010 Census Evaluation Program and Early Results

✦Jennifer W Reichert, U.S. Census Bureau, 4600 Silver Hill Road, Room 4H465, Washington, DC 20233, *jennifer.w.reichert@census.gov*

Key Words: 2010 Census, 2020 Census, Evaluations

The U.S. Census Bureau has a long tradition of rigorously examining conduct and results of its decennial census. The 2010 Census Program for Evaluations and Experiments (CPEX) continues that tradition with over 20 evaluations aimed at analyzing various aspects of the many methods and operations deployed to complete the count of all U.S. residents. The evaluations for the 2010 Census cover six main topic areas: coverage measurement, coverage improvement, field operations, language program, questionnaire content, and marketing and publicity. To conduct the many evaluations for the 2010 Census, the Census Bureau developed key research questions prior to the start of the census. The research questions were developed, vetted, and approved by both seasoned analysts and managers to ensure the evaluation program would yield results that would prove useful for both assessing the success of the 2010 Census as well as for guiding research and plans for the 2020 Census. This paper will provide a summary of the evaluation studies and associated research questions, and will also provide early results from some of those studies.

The 2010 Census Experiments

✦Joan Marie Hill, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, *joan.marie.hill@census.gov*

Key Words: 2010 Census, experiments

The U.S. Census Bureau conducted the 2010 Census Experiments Program in keeping with a long tradition of research aimed at improving quality and reducing costs. Since 1970, the Census Bureau has implemented an experimental program to evaluate a variety of alternative methodologies and questionnaire design strategies. The 2010 Alternative Questionnaire Experiment focused on improving the race and Hispanic origin questions but also included other treatments, such as alternative address collection for improved within-household coverage and a Census 2000 short form-style questionnaire, which evaluates the effect of design changes made in the previous decade. The following 2010 Census Experiments were also included in the Program: Deadline Messaging/Compressed Schedule, Confidentiality Notification, Nonresponse Followup Contact Strategy, and Paid Advertising Heavy Up. This paper will provide the design of the 2010 experiments and early research results.

Early Proposed 2020 Census Research

✦Kevin Deardorff, U.S. Census Bureau, , *kevin.e.deardorff@census.gov*; Heather Madray, U.S. Census Bureau; Melissa Therrien, U.S. Census Bureau

Key Words: Census, 2020 Research, Self-response, Internet, Evaluations

The 2010 Census was the costliest Census to date, with a 63% increase in the cost per housing unit compared with 2000. Innovations are needed in order to prevent further large increases in cost in 2020. The Census Bureau has proposed six possible models for the 2020 Census, each of which offers new options for self-response and streamlines Census operations. This paper will discuss these six options and the early decade research that will provide the framework for decision-making in choosing the modes and methods for the next Census. This early decade testing will include extensive research on response modes, such as Internet, phone, and other electronic media to increase self-response; options for streamlining workload management and other field operations; and ways to improve address frame updating. The current 2010 Experiments, Evaluations, and Assessments program (EEA) will also

153 Advances in modeling non-traditional network data ●

General Methodology

Monday, August 1, 10:30 a.m.–12:20 p.m.

Estimating Hidden Population Size Using Respondent-Driven Sampling Data

✦Mark S Handcock, University of California - Los Angeles, Department of Statistics, 8125 Mathematical Sciences Building, Los Angeles, CA 90095-1554, *handcock@ucla.edu*; Krista Jennifer Gile, University of Massachusetts

Key Words: Bayesian Statistics, survey sampling, network, epidemiology, social science, likelihood

The study of hard-to-reach or otherwise "hidden" populations presents many challenges to existing survey methodologies. Examples include injection drug users and unregulated workers. These populations are characterized by the difficulty in sampling from them using standard probability methods. Typically, a sampling frame for the target population is not available, and its members are rare or stigmatized in the larger population so that it is prohibitively expensive to contact them through the available frames. Hard-to-reach populations in the US and elsewhere are under-served by current sampling methodologies. Respondent-Driven sampling (RDS) is one approach to collect data from networked populations. Most analysis of RDS data has focused on estimating aggregate characteristics of the target population, such as disease prevalence. However, RDS is often conducted in settings where the population size is unknown and of great independent interest. We present an approach to estimating the size of a target population based on the data collected through RDS. This is joint work with Krista J. Gile (UMass-Amherst) and Corinne M. Mar (University of Washington).

Post-Stratification and Network Sampling

Rachel Schutt, Google Inc; ✦Andrew Gelman, Department of Statistics, Columbia University, New York, New York, NY 10027 USA, *gelman@stat.columbia.edu*; Tyler McCormick, Columbia University

Key Words: Hierarchical modeling, post-stratification, network sampling

We propose a method for adjusting for bias in samples from networks using Bayesian hierarchical models. Our method combines previous work in post-stratification for standard surveys with recent work on network sampling and indirectly observed network data. A key feature of our approach is incorporating network structure into the hierarchical model to reduce bias in population or subpopulation-level estimates. We demonstrate our general framework using a sample of 500 men who have sex with men recruited using Respondent Driven Sampling in Buenos Aires, Argentina.

Birds of a Feather Shop Together: Predicting Adoption with Social Networks

✦Sharad Goel, Yahoo Research, , *goel@yahoo-inc.com*

Key Words: networks, sociology, marketing, prediction

Adoption is often predicted using individual-level attributes such as age, sex, and geographic location. The principle of homophily suggests that social data (e.g., the attributes of people with whom one is in contact) might also have predictive value, however a lack of such social network data has limited research into this question. To assess whether social data can improve predictive models in a variety of domains, we construct a network from email and instant message exchanges and apply it towards making individual-level predictions of retail spending, joining a recreational league, and reacting to online advertisements. In each instance, we find that network data help identify those individuals most likely to adopt, and moreover, that these data often improve upon traditional indicators. This work is joint with Daniel Goldstein.

154 JASA, Theory and Methods Invited Session, Adaptive Confidence Intervals for the Test Error in Classification

JASA, Theory and Methods

Monday, August 1, 10:30 a.m.–12:20 p.m.

Adaptive Confidence Intervals for the Test Error in Classification

✦Eric B Laber, University of Michigan Department of Statistics, , *laber@umich.edu*; ✦Susan A Murphy, University of Michigan Department of Statistics, , *samurphy@umich.edu*

The estimated test error of a learned classifier is the most commonly reported measure of classifier performance. However, constructing a high quality point estimator of the test error has proved to be very difficult. Furthermore, common interval estimators (e.g. confidence intervals) are based on the point estimator of the test error and thus inherit all the difficulties associated with the point estimation problem. As a result, these confidence intervals do not reliably deliver nominal cover-

age. In contrast we directly construct the confidence interval by use of smooth data-dependent upper and lower bounds on the test error. We prove that for linear classifiers, the proposed confidence interval automatically adapts to the non-smoothness of the test error, is consistent under fixed and local alternatives, and does not require that the Bayes classifier be linear. Moreover, the method provides nominal coverage on a suite of test problems using a range of classification algorithms and sample sizes

155 Recent advances in the applications and theory of spline smoothing ■●

International Chinese Statistical Association, Section on Nonparametric Statistics

Monday, August 1, 10:30 a.m.–12:20 p.m.

On Equivalent Kernels For (Periodic) Spline Estimators

✦Tatyana Krivobokova, Georg-August-Universitaet Goettingen, Goettingen, 37073 Germany, *tkrivob@gwdg.de*; Katja Schwarz, Georg-August-Universitaet Goettingen

Key Words: Penalized splines, Equivalent kernels

In this work we study general spline based estimators, where both the number of knots and the smoothing parameter control the (asymptotic) properties of estimators. This includes both extreme cases: least squares (regression) spline estimators (with smoothing parameter equal to zero) and smoothing spline estimators (with the number of knots equal to the number of observations). Applying Fourier techniques we study in a unified framework the asymptotic properties of spline based estimators. We discuss two asymptotic scenarios possible, depending on the number of knots chosen. In particular, we obtain the equivalent and reproducing kernel as a function of number of knots and smoothing parameter and present some results on the local asymptotics for these general spline estimators.

Generalized Additive Modelling Of Credit Rating

SHUZHUAN ZHENG, Department of Statistics and Probability, Michigan State University; ✦Rong Liu, University of Toledo, OH 43606 USA, *rong.liu@utoledo.edu*; Lijian Yang, Michigan State University; Lifeng Wang, Michigan State Unvieristy

Key Words: Generalized Additive Model, Confidence Band, Credit Rating, Accuracy Ratio, Spline, Kernel

One central field of modern financial risk management is corporate credit rating in which default prediction plays a vital role. Parametric models of default prediction lack flexibility of model specification that leads to a low prediction power, while nonparametric models with higher prediction power are computationally demanding. We propose spline-backfitted kernel (SBK) estimator in the context of generalized additive model (GAM) with simultaneous confidence bands and BIC constructed for components testing and selection. First, GAM performs well in dimension reduction that allows to deal with a large set of covariates chosen from financial statements. Second, SBK estimator is much more computationally expedient than kernel smoothing, thus very practical for the fast prediction, and inference can be made on component functions with confidence. Third, we developed a BIC criterion to search significant covariates in the GAM modelling situation. Our method is applied to predict default probability of 3,472 listed companies of Japan, and the prediction displays nearly perfect cumulative accuracy profile (CAP) curves and very high accuracy ratio (AR) scores.

Adaptive Smoothing With A Cauchy Process Prior

✦Paul Speckman, University of Missouri-Columbia, 134B Middlebush Bldg., Columbia, MO 65211-6100, *speckmanp@missouri.edu*

Key Words: smoothing spline, adaptive estimate, Cauchy process

The equivalence between spline smoothing, which penalizes the L_2 norm of a derivative, and Bayesian inference with an integrated Brownian motion prior is well known. In this talk, we explore the use of Cauchy process in place of Gaussian process priors. For simplicity, we use a discrete approximation to the derivative, so that an exact, explicit solution exists. We demonstrate by example that the resulting Bayes estimator has many of the desirable properties of a penalized estimator with an L_1 penalty, for example. Moreover, fully Bayesian inference is easily obtained through efficient MCMC techniques. The prior can be extended to provide adaptive smoothing on a lattice in two dimensions. Illustrations with simulated and real data are provided.

Spline Confidence Bands For Functional Derivatives

Guanqun Cao, Michigan State University; ✦Jing Wang, University of Illinois at Chicago, 851 S Morgan St (MC 249), Chicago, IL 60607, *wangjing@math.uic.edu*; Li Wang, University of Georgia; David Todem, Michigan State University

Key Words: B-spline, confidence band, functional data, derivative, semiparametric efficiency

This paper considers the problem of estimating the derivatives of the mean curve of dense functional data. In particular, confidence bands for the derivative curves are developed using polynomial B-splines. Both the spline estimator and its accompanying confidence band are shown to have the semiparametric efficiency in the sense that they are asymptotically the same as if all random trajectories are observed entirely and without errors. The confidence band procedure is illustrated through several numerical studies.

P-Splines Regression Smoothing And Variable Selection

✦Irene Gijbels, Katholieke Universiteit Leuven, Celestijnenlaan 200B, Box 2400, B-3001 Leuven (Heverlee), International B-3001 Belgium, *Irene.Gijbels@wis.kuleuven.be*

Key Words: P-splines regression, variable selection, regression models

In this talk we focus on penalized estimation in additive models and varying coefficient models. A main interest is also on variable selection for such models. Particular attention goes to recent variable selection procedures such as grouped Lasso, grouped SCAD, COSSO, but also to the nonnegative garrote method introduced originally for variable

selection in a multiple linear regression model. We show how the latter method combined with P-splines estimation leads to an estimation and variable consistent method in both the settings of additive models and varying coefficient models. The performances of this and other related selection procedures are investigated in a simulation study and illustrations on real data examples are provided. In the additive varying coefficient model a uniform framework is presented and the various grouped regularization procedures are studied. A discussion on implementation issues and computational algorithms is provided, and a comparative simulation study is given. This talk is based on joint works with Anestis Antoniadis, Sophie Lambert-Lacroix and Anneleen Verhasselt.

156 The Census and Statistical Policy Making: Lessons from History ■●

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

Monday, August 1, 10:30 a.m.–12:20 p.m.

Who Writes Census History

✦Margo Anderson, University of Wisconsin - Milwaukee, Dept of History, UW-Milwaukee, Milwaukee, WI 53201 USA, *margo@uwm.edu*

Key Words: Census, statistical policy, methodology, Historical research

Historians call it "the usable past." The paper will discuss: a) who writes census history and why, b) where and if it is disseminated (including in private papers, in census volumes, in special volumes published by the census bureau, online dissemination, and/or other books and reports published external to the census bureau), c) how such history is supported, sponsored or commissioned, and d) the uses of the history for census takers and users, and students of the history of statistics and the nation.

Replacing Austin: A Study Of Leadership Change At The Us Census Bureau

✦William Seltzer, Fordham University, Department of Sociology and Anthropology, 411 East Fordham Road, Bronx, NY 10458 USA, *seltzer@fordham.edu*

Key Words: Census history, confidentiality, ethics, patronage, sampling, statistical administration

Shortly after his first inauguration in 1933 President Roosevelt nominated William Austin to be Director of the Census Bureau. Austin, a Mississippi Democrat, was then an old Census hand having been first appointed to the Bureau in 1900. During the run up to the 1940 Population Census Austin evidently lost the confidence of the President due to his opposition to efforts by the FBI and the military intelligence agencies to gain access to confidential census information and to Austin's unwillingness to follow political guidance on patronage appointments. The patronage issue was particularly sensitive as the President grappled for control of the Democratic Party in anticipation of the 1940 elections. After Roosevelt won his third term, Austin was forced out as Director. His replacement was JC Capt a New Deal political functionary from Texas, described in one White House memorandum as having "no professional background or standing in his profession." The paper describes these events in detail, discusses their possible impact on subsequent Bureau actions, and considers their relevance to the current legislative effort to reform of the appointment process of the Census Bureau Director.

Statistical DÈj‡ Vu: The National Data Center Proposal of 1965 and Its Descendants

✦Rebecca S. Kraus, U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20746, *rebecca.s.kraus@census.gov*

Key Words: Census Bureau, data center, data sharing, history, Bureau of the Budget, government statistics

Issues concerning sharing of statistical information, linking data sets, and storing and preserving data collected by the federal statistical agencies have long sparked debate. This paper focuses on the National Data Center proposal of 1965, ensuing public concern over its privacy implications, and the response of the Bureau of the Budget and the U.S. Census Bureau. The purpose of this study is to identify the issues leading to the development of the proposal, as well as the consequences of the proposal, in order to inform current policy decisions, particularly in regard to the U.S. Census Bureau. Examples of subsequent efforts at statistical consolidation and data sharing highlight the persistent theme of statistical dÈj‡ vu.

Delivering What Users Want: The Evolution Of Census Bureau Small Area Data

✦Michael S. Snow, U.S. Census Bureau, 3K412E, 4600 SILVER HILL RD, Washington, DC 20233, *michael.s.snow@census.gov*

Key Words: small area data, census tract, census blocks, decennial census, history, sampling

Increasing demand for small area data has driven the evolution of the decennial census since the late nineteenth century. Responding to public health officials' need for data on relatively homogeneous units, the Census Bureau began tabulating data on subdivisions of a few cities in the 1890s. When social workers and business organizations joined public health officers in asking the Census Bureau for such data, the agency agreed to publish data based on their delineation of census tracts. To meet growing demand from marketers and government planners, the Census Bureau added data on census blocks in 1940 and later on census county divisions and other small areas. Since the 1970s, the need for small area data for legislative redistricting and transportation planning pushed the agency to extend nationwide the areas for which it provided small area data. The most recent evolution has arisen out of calls for more timely data. The American Community Survey in 2010 replaced the decennial long-form and began delivering small area data more than once a decade.

157 Statistics and Epidemiology in Aging and Dementia

Section on Statistics in Epidemiology, Section on Health Policy Statistics, Scientific and Public Affairs Advisory Committee

Monday, August 1, 10:30 a.m.–12:20 p.m.

Effects Of Selection Bias And Competing Risks On Factors Determining Conversions To Cognitive Impairments And Mixed Dementias

✦Richard J. Kryscio, University of Kentucky, Center on Aging, 800 South Limestone Street, Lexington, KY 40536, *kryscio@email.uky.edu*; Yushun Lin, University of Kentucky; Liou Xu, University of Kentucky; Erin Abner, University of Kentucky

Key Words: transition models, Markov, joint models, competing risks, selection bias, mixed dementias

We present an overview of research involving data from serial cognitive assessments of elderly individuals followed to dementia or death. Assessments are categorized into four states: intact cognition, mild cognitive impairment, global impairments, and clinical dementia, all of which are interval censored events. Clinical dementia comprises heterogeneous pathologies; underlying disease(s) may not be clarified until autopsy. The analytic goal is to define relationships among risk factors, transitions between cognitive states, and probability of dementia before death. Statistical issues include selection bias, competing risk of death, and the effect of mixed dementias. Mixed dementia analysis will focus on hippocampal sclerosis, which occurs frequently in the very old. Results based on transition models using either a nonstationary Markov chain with absorbing states including interval censored deaths or joint models for the actual residual lifetime and the cognitive states will be summarized. Applications to two mature cohorts will illustrate the results: Biologically Resilient Adults in Neurological Studies (BRAINS), an observational cohort, and the Nun Study, a population cohort.

Choice Of Time Scale In Analyzing Longitudinal Data On Cognition And Its Effect On Inference

✦Lei Yu, Rush Alzheimer's Disease Center, , *lei_yu@rush.edu*; Eisuke Segawa, Rush Alzheimer's Disease Center; Sue Leurgans, Rush Alzheimer's Disease Center; David A Bennett, Rush Alzheimer's Disease Center

Key Words: Longitudinal data, Cognitive function, Time scale

In longitudinal studies of cognition, investigators examine the long term trajectories of cognitive functions and explore risk factors for change. Similar to time-to-event analysis, one practical issue of modeling continuous cognitive outcomes is the choice of an appropriate time scale. On one hand, time varying age is by far the most important risk factor and potential confounding variable in the study of aging, and it emerges as a natural conceptual time scale; on the other hand, time-on-study with adjustment for age at baseline allows the modeling of heterogeneity within the cohort, and therefore is often chosen when analyzing longitudinal cohort data. We aim to investigate whether these time scales present comparable characterizations of change in cognition. Using simulated data and examples from the Religious Order Study, an ongoing longitudinal study of aging and dementia, we will compare the statistical inference from the two approaches by assuming three different scenarios (1) linear trajectories; (2) linear trajectories with cohort effects; and (3) terminal decline. This study is supported by grants P30AG10161 and R01AG15819 from the National Institute on Aging.

A Multi State Parametric Approach For Modeling Cognitive Dynamics

✦Arnold Mitnitski, Dalhousie University, 229-5790 University Ave, West Annex, 2d Floor, Halifax, NS B3H 1V7 Canada, *arnold.mitnitski@dal.ca*; Nader Fallah, Dalhousie University; Charmaine B Dean, Simon Fraser University; Kenneth Rockwood, Dalhousie University

Key Words: multi-state model, stochastic process, truncated Poisson distribution, mortality, cognitive decline, cognitive improvement

We present a novel parametric multi-state stochastic model to describe longitudinal changes in cognitive tests. Scores are modeled as a truncated Poisson distribution, conditional on survival to a fixed endpoint; the Poisson mean depends on the baseline score and covariates. The basic model has four parameters. Two represent the cognitive transitions of individuals with no cognitive errors at baseline ("cognitively fittest") and their survival. The two other parameters represent corresponding increments in the probability of transitions and of mortality as a function of baseline cognition. Note that this approach permits estimation of the probabilities of transitions in different directions: improvement, decline and death. It allows gradual changes to be discerned, before progression to diagnostic and staging thresholds that are often irreversible. The model does not suffer from the common limitation of being applicable to only a small number of states. This is illustrated by comparing the model's performance with polytomous logistic regression. The model's performance is illustrated by applying it to different cognitive measures (MMSE, CASI, ADAS-Cog) in several databases.

Methods For Modeling The Association Of Change Points With Risk Factors For Cognitive Decline In The Elderly.

✦Charles B. Hall, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, *charles.hall@einstein.yu.edu*

Key Words: change point, longitudinal data, profile likelihood, Bayesian, Alzheimer's disease, dementia

Cognition in the elderly often is relatively stable up to some point at which individuals at risk for mild cognitive impairment or dementia begin to experience accelerated decline. Change point models are useful for modeling this transition, as they directly estimate the rates of change before and after the acceleration. The dependence of the change point, and the rates of decline, may depend on risk factors. This association may be modeled in three ways: (1) A profile likelihood method where the rates of decline conditional on the change point are modeled using standard mixed linear model software, (2) a maximum likelihood approach where the change point and the rates of decline are simultaneously estimated using a nonlinear mixed effects model, and (3) a Bayesian approach using Markov Chain Monte Carlo simulation. The Bayesian approach has the advantage of the ability to model heterogeneity in the change point across individuals beyond that captured in the known, measured risk factors. The methods will be compared using data from the Bronx Aging Study to estimate the effect of a purported marker for cognitive reserve.

Optimum Designs For Clinical Trials To Test Disease-Modifying Agents On Alzheimer'S Disease

✦Chengjie Xiong, Department of Biostatistics, Washington University, 660 S Euclid Ave., Box 8067, St. Louis, MO 63110, *chengjie@wubios.wustl.edu*

Key Words: disease-modifying trials, optimum design, linear mixed models, intersection-union test

Therapeutic trials of disease-modifying agents on Alzheimer's disease (AD) differ from the symptomatic trials because the former require novel designs and analyses involving switch of treatments for at least a portion of subjects enrolled. Randomized start and randomized withdrawal designs are two examples. Crucial design parameters such as sample size allocations and treatment switch time are important to understand in designing such clinical trials. A general linear mixed effects model is proposed to formulate the appropriate hypothesis for the test of disease-modifying efficacy of novel therapeutic agents on AD. This model incorporates the potential correlation on the rate of cognitive or biomarker change before and after the treatment switch. Optimum sample size allocations and treatment switch time of such trials are determined according to the model. An intersection-union test through an optimally chosen test statistic is used for the test of treatment efficacy. Using reported data on symptomatic trials on AD, the proposed methodology is applied to design future disease-modifying trials on AD.

158 Multi-mode Surveys: Design and Effects ■

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

Monday, August 1, 10:30 a.m.–12:20 p.m.

Distinguishing Mode Selection Effects From Mode Response Effects In The Consumer Assessments Of Healthcare Providers And Systems (Cahps) Survey

✦Alan M. Zaslavsky, Harvard Medical School, Department of Health Care Policy, Boston, MA 021115 USA, *zaslavsk@hcp.med.harvard.edu*

Key Words: principal stratification, mixed mode survey, health care surveys

The Consumer Assessments of Healthcare Providers and Systems (CAHPS) surveys of Medicare beneficiaries are conducted initially by mailout-mailback of a paper instrument, with telephone followup of nonrespondents. Differences between mean responses by mail and telephone can occur due to a combination of mode selection (those responding by telephone differ from those responding by mail) and mode response effects (responses by telephone differ from those that would be obtained from the same respondents by mail). We identified these effects using an embedded experiment in which a random subsample of beneficiaries was approached first by telephone and afterwards by mail. Our analysis was conducted in a principle stratification framework, decomposing the universe of potential respondents into mail-only respondents, phone-only respondents, and those who would respond by either mode.

Framing Inference From Mixed Mode Surveys Using Causal Inference Framework

✦Trivellore Raghunathan, University of Michigan, , *teraghu@umich.edu*

Key Words: Bayesian Inference, Potential Outcomes, Nonresponse, Measurement Error, Panel Surveys, Mixed Modes

To improve coverage and response rates, many surveys employ designs where the survey instruments are administered using a mixed modes such as Mail, Web, Telephone and In-person. In some designs, the modes are changed as a part of refusal conversion process. What is the appropriate approach to analyze data from such mixed mode designs? One may decide to ignore the modes, if the measurement properties are similar across the modes but, what if there are mode differences? How should one construct inferences for the population quantities based on the data collected from such survey data? This paper proposes a causal inference framework where we generate potential populations under each mode and then combine these potential populations to form a single inference. A Bayesian framework is used develop inferences and thus fully account for differences in the mode effects.The data from a longitudinal survey that used single mode design in first wave and mixed mode design in subsequent two waves will be used to illustrate the methodology. Data will be useful to highlight issues where modes may affect differentially across variables. The methodology is evaluated using a simulation study.

Can We Drive Respondents To The Web? Experiemntalttests Of Multi-Mode Survey Implementation And Varied Survey Benefit Appeals

✦Danna Moore, Washington State University, , *moored@wsu.edu*

Key Words: mult-mode surveys, experiments survey design, benefit appeals, agriculture

Two primary objectives of this research are:1)to evaluate if respondents can be motivated to complete the survey in any mode (web, mail, telephone), and 2) can we further motivate completion by the lowest cost strategy-the web questionnaire. An experimental frame work was used to evaluate the role and effectiveness of mixed mode survey implementation in combination with other survey strategies that can impact response. While web based surveys have become increasingly more common in usage and in research, this methodology has not been thoroughly investigated for agricultural populations and for USDA sponsored surveys. The other strategies included variation of benefit appeals used in letters, length of questionnaire, and visual design of question presentation screens on the web. This study also demonstrates the constraint of keeping question presentation similar across modes and evaluates the visual design consideration of one question per web screen versus multi questions per web screen on respondent completion and item non-response. Experiments were carried out on an initial sample of 13,000 and were evaluated through multivariate analysis of sample variables.

Piggyback Survey Respondents And Mode: Lessons Learned From Design And Operations

✦Brad Edwards, Westat, 1600 Research Blvd., Rockville, MD 20009, *bradedwards@westat.com*

Key Words: survey design, multi-mode, piggyback, sample design

"Piggyback" surveys have at least two parts: data collected from an initial sample are used to spawn another sample. Designing the second sample so that it "piggybacks" on the first is typically much more efficient than other sampling approaches. For many piggyback surveys, data from the two parts are collected in different modes (e.g., face-to-face for one, telephone for the other). This presents a number of design and operational challenges. The same questions may be asked in both surveys; they may be designed in one mode, then adapted for the other mode without considering mode effects. Elapsed time between the specific respondent's interview in Sample A and the interview with the Sample B respondent he or she identified may also be a concern. If too much time elapses, the link between the two may be broken. The paper will discuss lessons learned from a number of experiences with these types of multi-mode surveys: child/parent or care provider; medical setting and staff members in the setting; and disabled household members and their caregivers.

Using A Multi-Mode Survey Design On A Panel Study Of New Businesses

✦David DesRoches, Mathematica Policy Research, 600 Alexander Park, Princeton, NJ 08648 USA, *ddesroches@mathematica-mpr.com*

Key Words: establishment surveys, entrepreneurship, multi-mode surveys, business surveys

Since 2005, Mathematica has conducted the Kauffman Firm Survey (KFS) for the Ewing Marion Kauffman Foundation. The baseline KFS survey recruited a panel of U.S. businesses which were founded in the same calendar year (2004) using a multi-mode web/CATI design. This group of businesses (the KFS panel) has been contacted annually for follow-up data collection since 2006. The main goal of the KFS is to investigate how new businesses are structured and funded in their early years, and to measure the changes in business financing and productivity over this period. With the use of this multi-mode design, much of the data collection in the follow-surveys has been migrated from CATI to the web, reducing the costs and respondent burden associated with extensive telephone follow-up efforts. This paper will explore the experiences of recruiting a panel of establishments through a multi-mode survey, as well as the technological improvements made over the course of the study. The paper will also examine the cost effects of increasing web survey data collection in the follow-up surveys.

159 Advances in Adaptive Clinical Trials

Biopharmaceutical Section, Biometrics Section, ENAR
Monday, August 1, 10:30 a.m.–12:20 p.m.

Monitoring Adaptive Clinical Trials; Case Studies

✦Zoran Antonijevic, Center for Statistics in Drug Development Innovation, Quintiles, 8 Teahouse Ct., Durham, NC 27707, *Zoran.Antonijevic@Quintiles.com*

Key Words: adaptive design, clinical trial, interim decision process, interim monitoring

The emphasis of this presentation will be on the adaptive clinical trials interim decision process. Three case studies will be presented. For each case study the design will first be described, which will then be followed by description of the data monitoring model. First case study has a model in which all information is communicated to the sponsor by the independent statistical center (ISC). In second case study all information goes from the ISC to the DSMB, which then makes a recommendation to the sponsor. In third case study efficacy decision is communicated to the sponsor directly by the ISC, while safety recommendation is presented by the DSMB.

Evaluating The Usefulness Of Adaptive Designs In Practice: Going Beyond Statistical Characteristics

Jose Carlos Pinheiro, Johnson & Johnson PRD; ✦Chyi-Hung Hsu, Johnson & Johnson PRD, , *chsu3@its.jnj.com*

Key Words: adaptive design, clinical trial, dose-ranging study

When evaluating the appropriateness an adaptive design in drug development practice, one needs to take into account not only its statistical operating characteristics, such as power, precision of estimates, and sample size, but also its operational and implementation requirements and potential hurdles, such as drug supply and patient recruitment. In this talk, we use the planning of an adaptive dose-ranging study in a neuro-degenerative indication to motivate and illustrate the different factors that need to be taken into account in the design and evaluation of adaptive designs. A multi-disciplinary group was involved in the planning of the study, which was critical to allow an integrated evaluation of the potential implementation hurdles and benefits of the proposed design.

Design And Operational Issues In Clinical Trials With Adaptive Design

✦Gang Chen, Johnson & Johnson, 920 Route 202, Raritan, NJ 08502, *GChen11@its.jnj.com*

Key Words: adaptive design, clinical trial, type I error

The use of adaptive design in clinical trials is often attractive to pharmaceutical industry because this type of design may be more efficient in collecting information and may increase the likelihood of success. However, there are two major issues in such a design and trial conduct: 1) how do we avoid potential "bias" due to the operation; 2) How do we control "type I error" since some factor may not be completely "prespecified". In this talk both issues will be discussed and an example (trial design) will be given to illustrate those issues.

Adaptive Design: Some Theories And Examples

✦Xiaolong Luo, Celgene Corporation, 86 Morris Avenue, Summit, NJ , *xluo@celgene.com*

Key Words: adaptive design, clinical trial, type 1 error

While the concept of adaptive design for clinical trials has become well known through many recent publications and regulatory guidance, what is less recognized is its evolutional impact on general statistical framework and methodology, which attributes to some well known controversies in its applications. In this talk, we will explain how the commonly used statistics methodology is built on i.i.d. observations and why that assumption may not be sufficient for many clinical trials. We will introduce our recently developed statistical method that overcomes the typical i.i.d. limitation and naturally fits with adaptive design. We will use numerical examples and a completed adaptive clinical trial to evaluate the performance of the method in traditional criteria such as bias and type 1 error rate.

Adaptive Design With Some Practical Issues

✦Bo Yang, Merck, 2015 Galloping Hill Rd, Kenilworth, NJ 07033, *bo.yang2@merck.com*

Key Words: adaptive design, clinical trial, interim monitoring

In recent years, adaptive design methods have attracted much attention in clinical research due to its flexibility and efficiency. If use properly, adaptive approaches could facilitate early identification of therapeutic effects, conserve patient resources for treatments that have a higher probability of success or minimize patient exposure to ineffective treatments. However, adaptive designs are not without issues. Such designs may be more prone to bias in the interpretation of the results due to close monitoring and changes implemented based on the interim results. In this talk, we will review some of the practical issues that are commonly encountered when implementing adaptive design methods in clinical trials.

160 Experimental Design and the Modern Economy ■●

Business and Economic Statistics Section, International Indian Statistical Association, Scientific and Public Affairs Advisory Committee

Monday, August 1, 10:30 a.m.–12:20 p.m.

Do You Know Or Do You Think You Know? Creating A Testing Culture At State Farm

✦Andrew Pulkstenis, State Farm Insurance, 1512 Beaver Lake Drive, Mahomet, IL 61853, *akpulkst@aol.com*

Key Words: DOE, experiment, testing, business analytics, design

"Analytics" is the new buzzword in business, but even analytic firms are underusing or neglecting the powerful tool of designed experiments. I will share what I've been doing at State Farm over the past year to move our firm from minimal business experimentation to a more ambitious testing culture, as well as provide a step-by-step guide on how one can change or create a testing culture in their firm. I'll share a real example from State Farm's online presence, and close by discussing the comprehensive optimization effort this success has kicked off for 2011. The presentation will provide some tangible takeaways for the practitioners in the crowd who desire to improve their own strategies through designed experiments.

Testing At Capital One Auto Finance

✦Rose Brunner, Capital One Financial, , *Rose.Brunner@CapitalOne.com*; Leonard Roseman, Capital One Financial

The Capital One Auto Finance (COAF) Price Plus test aimed to reveal how three critical profitability factors drive dealer behavior in auto finance, for the purpose of expanding thin margins in this competitive arena. The test has several interesting features including a split plot design, unbalanced sampling due to cost constraints, power limitations and computations, concurrent testing, and early shut-down of some high-cost treatment combinations. In this session, we will outline the critical business dynamics and provide overview of how the technical issues were addressed.

Benefits & Challenges Of Experimental Design In The Chemical Industry

✦Stephanie Pickle DeHart, DuPont, 611 Crystal Anne Lane, Roanoke, VA 24019, *stephanie.p.dehart@usa.dupont.com*

Key Words: experimental design, DOE, chemical, science, industrial statistics

Design of experiments (DOE) is a powerful and cost effective technique that can provide data driven solutions to many problems. DOE has been successfully applied in a multitude of industrial applications including research and development, operations, and marketing and sales. We will discuss the benefits and challenges of using DOE in the chemical industry while highlighting some examples from DuPont, a global science company.

Experimental Design In The Modern Economy

✦Thomas J Kirchoff, Capital One Services, LLC, 15000 Capital One Drive, Richmond, VA 23238, *tom.kirchoff@capitalone.com*

Hierarchical designs, such as split plot designs, are often necessary in the banking and financial services sectors. For example, it may be possible to test some factors at the customer level, but other factors in the same test must be assigned at the bank branch level. A question of paramount interest in any financial services test is that of sample size. In this presenatation, two different methods for determining sample size at the "whole plot" level are discussed.

161 Bayesian Methods for Longitudinal Data Analysis ●

Section on Bayesian Statistical Science, ENAR

Monday, August 1, 10:30 a.m.–12:20 p.m.

Joint Bayesian Modeling Of Irregularly Measured Multivariate Longitudinal Nutrient Consumption And Growth Data

✦Sherry Lin, University of California Los Angeles, , *sherry1476@yahoo.com*; Robert E. Weiss, University of California, Los Angeles

Key Words: nutrient patterns, missing data, forward-backward recursion, MCMC, dynamic model

We have irregularly measured longitudinal nutrient intake data and asynchronously measured longitudinal growth data. The nutrient consumption data is high dimensional and exhibits multivariate correlation across different nutrients and serial correlation across time. A subject's recent dietary intake, along with his or her intake history, can both be associated with growth. We propose a joint model which uses a small number of dynamic latent factors to model the observed nutritional intake covariates and to predict the growth outcome. The latent factors are interpreted as underlying dietary intake patterns which evolve according to a continuous-time autoregressive process. The model simultaneously estimates the correlation structures of the growth and nutrient data. We formulate the model under a Bayesian framework, using a forward-backward algorithm to estimate the latent factors. The model is applied to nutrition data from a Kenya school intervention study used to analyze muscle development of children under different nutritional snack regimens.

A Hierarchical Bayesian Model For Multivariate Timeline Follow-Back Histories

✦Adam King, University of California, Los Angeles, 641 Gayley Ave., Apt. 102, Los Angeles, CA 90024, *aking@ucla.edu*; Robert E. Weiss, University of California, Los Angeles

Key Words: Bayesian, Hierarchical, Longitudinal, Multivariate

Timeline follow-back (TLFB) is a method for retrospectively eliciting histories of behaviors and circumstances from study subjects. For each such time-dependent trait of the subject, time segments on which that trait was constant are recorded along with covariates describing that trait on that time segment. We propose a hierarchical Bayesian model for this data structure, along with supporting computational tools and graphics. We apply these methods to lifetime TLFB histories of illicit drug use and related traits and experiences of 508 subjects.

Continuous Trajectory Modeling When Data Are Collected Longitudinally At Common Discrete Time Points

✦John Boscardin, San Francisco VA and Univ of Cal San Francisco, 4150 Clement Street (181G), San Francisco, CA 94121 USA, *john.boscardin@ucsf.edu*

Key Words: mixed effects models, spline models, non-normal data distributions

Detailed modeling of continuous trajectories for longitudinal data is often hampered by a small number of discrete time points for data collection. Changing the time scale of the analysis can present a solution. We discuss two classes of examples: (i) using age of the subject instead of time on study, and (ii) centering the followup time around an event time for which exact date of occurrence is available (e.g. date of hospitalization). The examples will be discussed in the context of a longitudinal study on older adults where most measures are collected at regular bi-annual intervals but exact dates are available for a number of events of interest. Our interest lies in modeling pre- and post-event trajectories of a longitudinal measure. Informative dropout due to subject death is accounted for using a joint model of the longitudinal and mortality data.

Bayesian Longitudinal Models For Duel Variable And Covariance Selections

✦Xuefeng Liu, East Tennessee State University, Johnson City, TN 37614, *lix01@etsu.edu*

Key Words: Longitudinal Models, Stochastic Search Variable Selection, Cholesky Decomposition, MCMC, Data Augmentation

Bayesian models are proposed to integrate parsimonious covariance structure into stochastic variable selection for longitudinal binary data. The covariance matrix is re-parameterized through Cholesky decomposition such that the Cholesky factor is likely to have close-to-zero elements. Hierarchical priors are used to identify the ``zero' elements in the Cholesky factor and the removable covariates in the models. variable selection techniques are used to parsimoniously model these components. A computationally efficient MCMC sampling algorithm is developed for sampling from the posterior distribution, using data augmentation steps to handle missing data. Several technical issues are addressed to implement the MCMC algorithm efficiently. The models are applied to a longitudinal study on aging.

162 Savage Award Session ●

Section on Bayesian Statistical Science, Section on Statistical Computing

Monday, August 1, 10:30 a.m.–12:20 p.m.

Adaptive Sequential Monte Carlo Methods

✦Julien Cornebise, University College London, International UK, *julien@stats.ucl.ac.uk*; Eric Moulines, Telecom ParisTech; Eric Moulines, Telecom ParisTech

Key Words: Sequential Monte Carlo algorithms, Adaptation, Kullback Leibler divergence, Computational statistics, Central Limit Theorems, Particle filter

We focus on the design and analysis of Adaptive Sequential Monte Carlo (SMC) algorithms. Our aim: enable them to automatically tune their parameters -- such as first stage resampling weights and proposal kernel -- for maximum computational efficiency and accuracy of the resulting estimates. We first formalize and study the existing practices from a theoretical point of view, asymptotically linking the coefficient of variation and the entropy of the importance weights (currently used on an empirical basis) to chi-square and Kullback-Leibler divergences (KLD) between distributions on an extended space. We develop new criteria decoupling of the adaptation of the first stage weights and that of the proposal kernel. Based on those and with inspiration from Stochastic Approximation and Monte Carlo EM, we build new algorithms able to deal with intricate non-linearities and multi-modality, and illustrate their performances in terms of KLD reduction and distribution of importance weights on several thoroughly examined numerical examples.

Adaptive Monte Carlo For Bayesian Variable Selection In Regression Models

✦Demetris Lamnisos, Cyprus University of Technology, 31 Archbishop Kyprianos Str., 50329, 3603, Limassol, , *demetris.lamnisos@cut.ac.cy*

Key Words: Linear regression, Probit regression, Metropolis-within-Gibbs

The availability of datasets with large number of variables has lead to interest in the use of variable selection methods for regression models with many regressors. In this work, we concentrate on Bayesian variable selection methods where efficient Markov chain Monte Carlo methods need to be developed to explore the vast model space. A Metropolis-Hastings sampler is implemented with a model proposal that generates a candidate model by randomly changing components of the current model. This is similar to a Random Walk Metropolis (RWM) sampler, which proposes a new state as perturbed version of the current state. We generalize this model proposal to include a tuning parameter that determines the degree of "localness" for the sampler and behaves similarly to the scale parameter of the RWM. The application of this proposal to datasets with many variables suggests that the optimal sampler occurs for a parameter which leads to an average acceptance rate close to 0.234. Therefore, we develop an adaptive sampler that automatically adapt this tuning parameter and allows efficient computation in these problems. The method is applied to examples from normal linear and probit regression.

Bayesian Policy Support: Using Computer Models To Develop Adaptive Strategies

✦Daniel Williamson, Durham University, Durham, International United Kingdom, *daniel.williamson@durham.ac.uk*

Key Words: Policy Support, Computer Models, Emulation, Reification, Dynamic Programming

Many of the world's toughest policy decisions rely on the information from large computer models of complex physical systems. In addressing climate change, for example, policy makers are informed by climate projections using the worlds climate simulators. The question of how the uncertainties in the models and the system should be synthesised in order to provide decision support has, thus far, been tackled using simplified models of the system, concentrating on only a handful of possible policies and often neglecting to account for large sources of uncertainty. We present a decision theoretic approach that treats all possible policies simultaneously, accounts for the possibility of future observation of the system and of improved computer models to help revise policy, makes use of the best system models available, and includes major sources of uncertainty such as model discrepancy. The decision tree defined by this approach has an infinite number of branches at each point and cannot be solved using standard methods. We introduce a Bayesian approach to providing decision support for this problem that is based on current methods in the design and analysis of computer experiments.

Adaptive Error Modelling In Mcmc Sampling For Large Scale Inverse Problems

✦Tiangang Cui, The University of Auckland, Level 3, Uniservices House, 70 Symonds Street, Auckland, International 1142 New Zealand, *tcui001@aucklanduni.ac.nz*; Colin Fox, University of Otago; Mike O'Sullivan, The University of Auckland

Key Words: adaptive MCMC, delayed acceptance, inverse problem, reduced order models, enhanced error model, geothermal reservoir modelling

We present a new adaptive delayed-acceptance Metropolis-Hastings (ADAMH) algorithm that adapts to the error in a reduced order model to enable efficient sampling from the posterior distribution arising in complex inverse problems. This use of adaptivity differs from existing algorithms that tune random walk proposals, though ADAMH also implements that. We build on the conditions given by Roberts and Rosenthal (2007) to give practical constructions that are provably convergent. The components are the delayed-acceptance MH of Christen and Fox (2005), the enhanced error model of Kaipio and Somersalo (2007), and adaptive MCMC (Haario et al., 2001; Roberts and Rosenthal, 2007). ADAMH is used to calibrate large scale numerical models of geothermal fields. It shows good computational and statistical efficiencies on measured data. We expect that ADAMH will allow significant improvement in computational efficiency when implementing sample-based inference in other large scale inverse problems.

The Generalized Multiset Sampler

✦Hang J Kim, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Cockins Hall Room 404, Columbus, OH 43210 United States, *kim.2243@osu.edu*; Steven N MacEachern, Ohio State University

Key Words: Multiset sampler, MCMC, Multimodal, Metropolis algorithm, Poor mixing, Mixture model

The multiset sampler (MSS) proposed by Leman et al. (2009) is a new MCMC algorithm, especially useful to draw samples from a multimodal distribution, and easy to implement. We generalize the algorithm by redefining the MSS with explicit description of the link between target distribution and sampling distribution. The generalized formulation makes the idea of multiset (or k-tuple) applicable not only to Metropolis-Hastings, but also to other sampling methods. The basic properties of sampling distribution are provided. Drawing on results from importance sampling, we also create effective estimators for both the basic multiset sampler and the generalization we propose. Simulation examples confirm that the generalized multiset sampler (GMSS) is a general and easy approach to deal with multimodality and to produce a chain that mixes well.

163 Clickers in Statistics Classes: Connecting Research and Practice ■●

Section on Statistical Education, Section on Quality and Productivity, Section on Teaching of Statistics in the Health Sciences

Monday, August 1, 10:30 a.m.–12:20 p.m.

Clickers In Statistics Classes: Connecting Research And Practice

✦Jennifer J Kaplan, Dept. of Statistics and Probability, Michigan State University, East Lansing, MI 48823 USA, *kaplan@stt.msu.edu*; ✦Josh J. Bernhard, Iowa State University, Department of Statistics, Ames, IA 50011-1210, *berni@iastate.edu*; ✦Patti B. Collings, Brigham Young University, 231 TMCB BYU, Provo, UT 84602, *collingsp@stat.byu.edu*; ✦Ulrike Genschel, Iowa State University, Department of Statistics, Ames, IA 50011-1210, *ulrike@iastate.edu*; ✦Herle McGowan, North Carolina State University, Department

of Statistics, 2311 Stinson Dr, Campus Box 8203, Raleigh, NC 27695-8203, *mcgowan@stat.ncsu.edu*

Key Words: Personal Response Systems, Research and Pedagogy, Classroom Technology

This panel is composed of statistics instructors who have used clickers in instruction and statistics education researchers who have studied the effectiveness of clickers on student outcome variables through specifically designed experiments. Examples of student outcome variables are learning, engagement, and student attitudes toward statistics. While the statistics education researchers will present primarily concepts and ideas related to the quantitative nature of their studies, the statistics instructors will discuss implementation decisions made based on their experiences and/or existing "best practices" literature, resulting from observational studies and qualitative feedback from students and instructors. The audience will have the opportunity to ask questions, and the session will provide a discussion about how the research results can lead to informed pedagogical decisions made by instructors. It is the goal of the panel to provide not only clicker oriented information and assessment tools that instructors can implement, but also to serve as an example of how quantitative research in general can and should affect future pedagogical practices for researchers and instructors.

164 Optimal Design of Experiments for Multiple Objectives ■●

Section on Quality and Productivity, International Indian Statistical Association, Section on Physical and Engineering Sciences

Monday, August 1, 10:30 a.m.–12:20 p.m.

Optimal Design of Experiments for Multiple Objectives

✦Timothy Robinson, University of Wyoming, , *TJRobin@uwyo.edu*; ✦Bradley Jones, SAS, SAS Campus Drive, Cary, NC 27513, *bradley.jones@jmp.com*; ✦Roselinde Kessels, University of Antwerp, , *roselinde.kessels@ua.ac.be*; ✦Chris Nachtsheim, University of Minnesota, , *nacht001@umn.edu*

Key Words: Design of Experiments, Optimal Designs, Response Surface, Choice Experiments, Robust Parameter Design

Box and Draper (1959) suggest a suite of criteria to evaluate when creating a designed experiment. This list of considerations suggests balancing estimation of model parameters and good prediction of new observations for an assumed model with other criteria including robustness to model misspecification and difficulties in collecting the data. For the practitioner it is often difficult to determine which subset of the list to highlight. Then, once the practitioner succeeds in prioritizing the list for the problem at hand, it is a formidable task to implement a design construction procedure that optimally addresses the multiple competing objectives. In this discussion, we present a number of the potential criteria to consider for different scenarios, as well as several different approaches to finding optimal designed experiments for multiple objectives. The panel will consider screening experiments, response surface designs, choice experiments and robust parameter designs through a series of examples, which outline the process for deciding which criteria to consider as well as how to implement computer algorithms for obtaining the optimal designs. We will compare the solutions w

165 Are survey statisticians prepared to address government needs for disability statistics? ■

Committee on Statistics and Disability, Statistics Without Borders, Scientific and Public Affairs Advisory Committee

Monday, August 1, 10:30 a.m.–12:20 p.m.

Are Survey Statisticians Prepared To Address Government Needs For Disability Statistics?

✦David W Keer, National Institute on Disability & Rehabilitation Research, 4711 Hollywood RD, College Park, MD 20740, *David.Keer@ed.gov*; ✦Jennifer Madans (invited), National Center for Health Statistics, , *jhm4@cdc.gov*; ✦John Hough (invited), National Center for Health Statistics, , *jph7@cdc.gov*

Key Words: Disability, International, human rights, surveys

On December 13, 2006, the first human rights treaty of the 21st century was adopted by the United Nations, the UN Convention on the Rights of Persons with Disabilities. As of today, 146 countries have now signed the Convention. The purpose of the UN Convention is to promote, protect and ensure the full and equal enjoyment of all human rights and fundamental freedoms by all persons with disabilities, and to promote respect for their inherent dignity. A strong Convention needs good statistics to monitor its course. How effective are survey statisticians going to be in meeting the needs of such a Convention? Article 31 of this Convention agrees that States Parties will undertake to collect and disseminate appropriate statistical information to enable governments to objectively formulate and monitor its implementation. Disability statistics produced by national statistical offices and researchers around the world have been notoriously plagued with problems of incomparability in definitions, concepts and standards. This session reviews recent achievements made by national statistical offices to improve disability statistics by setting the necessary standards and methods for their

166 Current issues in forensic science ■●

Section on Physical and Engineering Sciences, Committee of Representatives to AAAS

Monday, August 1, 10:30 a.m.–12:20 p.m.

Current Issues In Forensic Science

✦Constantine Gatsonis, Brown University, , *gatsonis@stat.brown.edu*; ✦Sarah Chu, Innocence Project, , *schu@innocenceproject.org*; ✦Clifford Spiegelman, Texas A&M University, , *cliff@stat.tamu.edu*

Key Words: forensic science, NRC, anthrax, toolmarks, legislation

This panel discusses several recent developments in forensic science. First, former NRC panelists discuss two recent NRC reports: a just released report on the 2001 anthrax attacks NRC (2011), and the recently released NRC report "Strengthening Forensic Science in the United States: A Path Forward" (2009) (which John Holdren, the President's Science Advisor, described as a "blockbuster"). Then, members of the

Innocence Project discuss proposed legislation responding to recommendations in the 2009 NRC report. Finally, a member of the statistical community discusses the state of firearm toolmark evidence in courts post the 2009 NRC report.

167 Bayesian Case Studies and Applications

Section on Bayesian Statistical Science

Monday, August 1, 10:30 a.m.–12:20 p.m.

Feedback And Modularization In A Bayesian Meta-Analysis Of Tree Traits Affecting Forest Dynamics

✦Kiona Ogle, Arizona State University, School of Life Sciences, PO Box 874501, Tempe, AZ 85287-4501, *Kiona.Ogle@asu.edu*; Jarrett Jay Barber, Arizona State University

Key Words: incomplete reporting, feedback control, modularization, ecological meta-analyis, hierarchical Bayesian, forest ecology

We describe a unique application of modularization, or 'feedback control', in the context of a Bayesian meta-analysis of literature information. Numerous missing data are common to meta-analyses, and, in this study, poor chain mixing and identifiability issues resulted. In response, we modularized model components such that missing covariate data do not allow feedback between modules to affect parameters in the covariate module (direct feedback control) or to affect covariate effects parameters in the mean model for the response (indirect feedback control). Our use of direct and indirect feedback control improved mixing and convergence, yielding realistic pseudo-posteriors. Such modularization addresses limitations of existing meta-analytic methods by accommodating incomplete reporting and by considering all quantities as stochastic, including sample means (response), sample sizes, standard errors, and covariates, reported or not. We illustrate our approach with literature information on specific leaf area, a key parameter in models of tree growth and forest dynamics. We discuss problems that arise from feedback between modules and provide ecological arguments for modularization.

Secure Bayesian Model Averaging for Horizontally Partitioned Data

✦Joyee Ghosh, University of Iowa, , *joyee-ghosh@uiowa.edu*; Jerome P. Reiter, Duke University

Key Words: Bayesian model averaging, Data confidentiality, Disclosure limitation, Markov chain Monte Carlo, Variable selection

The Bayesian paradigm allows one to incorporate covariate set uncertainty in linear regression and its generalizations by considering models corresponding to all possible combinations of covariates. Depending on the goals of the study, posterior probabilities of models can be used for model selection or model averaging. The setting in which multiple data owners or agencies possess data on different subjects but the same set of covariates is known as horizontally partitioned data. Such owners are often interested in global inference or prediction using Bayesian model averaging (BMA) for the combined data. However, sharing data across agencies may be infeasible for confidentiality issues. For linear regression, we introduce an approach called secure Bayesian model averaging (BMA), which performs exact BMA for the combined data without sharing information on individual subjects, using a technique called secure summation. For binary regression we first describe an exact approach to BMA, which requires several rounds of secure summation. As an alternative, we also suggest an approximate approach which requires only one round of secure summation, as in the case of linear regression.

A Bayesian Adjustment Of The Hp Mortality Law Using A Nonlinear Switching Regression Model

✦Dilli Bhatta, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, *drb122@wpi.edu*; Balgobin Nandram, Worcester Polytechnic Institute; Rong Wei, National Center for Health Statistics, CDC, USA

Key Words: Mortality, Helligman-Pollard, Switching Regression, Bayesian

Curve fitting of mortality for the US population is an important project at the National Center for Health Statistics (NCHS). We use the eight-parameter Heligman-Pollard (HP) empirical law to fit the curve. It consists of three nonlinear curves: child mortality curve, mid-life mortality curve and adult mortality curve added together. It is now well-known that the eight unknown parameters in the HP law are difficult to estimate. We consider a novel idea to fit the three curves (splines) separately, and then connect them smoothly at the two knots. Because these curves do not have turning points, to connect the curves smoothly we express uncertainty about the knots. A priori we have ranges of values for the knots. Thus, the Bayesian paradigm is particularly attractive. We discuss estimation of the mortality curves for US data, 1999-2001, and English and Welsh data, 1988-1992.

Comparison Between Bayesian And Frequentist Methods Using Winbugs And Sas For Two Group Proportions From A Comparative Study

✦Aijun Gao, i3 Statprobe, 9050 Centre Pointe Drive, #400, West Chester, OH 45069 USA, *aijun.gao@i3statprobe.com*; Kyoungah See, Lilly; Jing Zhang, Miami University

Key Words: Bayesian method, Frequentist method, Incidence rates, WinBUGS, GENMOD, MCMC

A comparative study included about 200 patients appropriate for a treatment therapy. The study assessed a categorical response about the treatment therapy from subjects using it in the community setting to self-administer study drug at baseline and two follow-up visits. There were about 100 patients in the current user group and patients in the non-current user group. Patient response incidence rates and the 95% confidence intervals were estimated using Frequentist methods (PROC FREQ and PROC GENMOD). As a comparison, the incidence rates were analyzed using Bayesian methods (WinBUGS, PROC GENMOD, and PROC MCMC). Different priors were tested to check the sensitivity. Similar results were obtained from Frequentist and Bayesian methods. The Bayesian approach provides immediate interpretation for estimated quantities and predictions and allows possible historical information to be included in studies. It also provides the posterior probability of any event, which makes the comparison of the incidence rates between current user group and non-current user group easier.

Bayesian Analysis And Applications Of Unified Competing Risks Cure Rate Models

✦Suchitrita Sarkar, Northern Illinois University, Du Sable 374, DeKalb, IL 60115, *sarkar@math.niu.edu*; Sanjib Basu, Northern Illinois University

Key Words: Survival Analysis, Bayesian Analysis, Cure Rate, Competing Risks

Important measures of recovery from cancer include survival rate or cure fraction. However, when the cause of death is one among several possible causes, it becomes difficult to get a precise estimate of the survival rate from cancer. The competing risks approach, which takes into account the risks occurring simultaneously from cancer and other causes, is appropriate here. When cure is possible, we see that the survival curve forms a plateau after sufficient follow-up. In such cases, the use of cure rate models is more realistic. In this article, we develop a Unified Competing Risks Cure Rate Model within the Yakovlev cure rate model framework. We describe Bayesian analysis of this model and discuss the conceptual and methodological issues related to model building and model selection. This model is generalized to a wider framework using latent factors; the activation of at least one of these factors is assumed to cause cancer. These models are motivated and illustrated by survival data on female breast cancer patients from the SEER program of the National Cancer Institute. We further compare the performance of the proposed models with other competing models in simulated data sets.

Bayesian Optimization In Scheduling Screening Tests

✦Yi Cheng, Indiana University South Bend, , *ycheng@iusb.edu*

Key Words: Bayesian,, Decision-theoretic approach, Optimal screening strategies, Utility, Early detection, Sensitivity

The goals of early detection of a particular disease with available screening test(s) are to reduce morbidity and mortality. Optimal screening strategies are expected to carefully balance these goals against the associated economic burden to patients and to health care systems. This research, using Bayesian decision-theoretic approach, investigates the problem of optimization by incorporating age-specific screening sensitivity, sojourn time, and incidence into a utility function.

Bayesian Hierarchical Monotone Regression I-Splines For Dose-Response Modeling And Drug-Drug Interaction Analysis: Application To In Vitro Studies

✦Violeta Hennessey, AMGEN, 1800 W. Hillcrest Dr. Apt 271, Newbury, CA 91320, *v.g.hennessey@gmail.com*; Veera Baladandayuthapani, UT MD Anderson Cancer Center; Gary Rosner , Johns Hopkins

Key Words: Median-effect principle, Combination Index, Emax model, Bayesian Effect Interaction Index, functional data analysis

We developed a flexible nonparametric method for meta-analysis of independently repeated dose-response experiments. Under the assumption of a non-increasing (or non-decreasing) monotone dose-response relationship, we make use of monotone regression I-splines (Ramsay, 1988) for estimating the mean dose-response curve. The complexities of monotonicity constraint can easily be accommodated under our Bayesian framework. We incorporate the splines into a Bayesian hierarchial model to address variability between-experiments, within-experiment (between-replicates), and variability in the controls. Markov chain Monte Carlo (MCMC), as implemented in WinBUGS, is used to fit the model to the data and carry out posterior inference on quantities of interest (e.g., inhibitory concentrations, Loewe Interaction Index for drug-drug interaction analysis). In addition, we explore a decision rule to assess drug-drug interaction. We compare our approach to analysis using the conventional parametric Median-Effect Principle/Combination Index method (Chou and Talalay, 1984).

168 Multivariate and Recurrent Survival

Biometrics Section, ENAR, WNAR

Monday, August 1, 10:30 a.m.–12:20 p.m.

A Chi-Squared Type Goodness Of Fit Test For Recurrent Event Data

✦Akim Adekpedjou, Missouri S & T, Rolla, MO 65409, *akima@mst.edu*; Gideon Zamba, Department of Biostatistics

Key Words: Recurrent Events, Gaussian Process, Pitman's Alternative, Goodness of Fit

Goodness of fit of the distribution function governing the time to occurrence of a recurrent event is considered. We develop a chi-squared type of test based on a nonparametric maximum likelihood estimator (NPMLE) for testing the distribution function of the inter event time of recurrent event data. The test is of the inter-event time distribution for recurrent events. The test compares a parametric null to the NPMLE over k partitions of a calendar time s. We investigate small sample and asymptotic properties of the test as well as power analysis against a sequence of Pitman's alternatives. Four variants of the test resulting from a combination of variance estimators and censoring were studied. The conclusion that transpires from the finite sample simulation study is that significant level is achieved when the right-censoring random variable is not ignored and $k > 3$. For exponential model, the tests are less powered to detect lighter right tail distribution than they are for left tails (contrary to Weibull model findings). The Weibull model has a slow response to heavier left tail distributions. We apply the test to a real-life recurrent event data.

Smoothing Spline Anova Frailty Model For Recurrent Event Data

✦Pang Du, Virginia Tech, Department of Statistics, 406-A Hutcheson Hall, Blacksburg, VA 24061, *pangdu@vt.edu*; Yihua Jiang, StubHub Inc.; Yuedong Wang, University of California at Santa Barbara

Key Words: Recurrent event data, Gap time hazard, Nonparametric hazard model, Frailty, Smoothing spline ANOVA, Stochastic approximation

Gap time hazard estimation is of particular interest in recurrent event data. This paper proposes a fully nonparametric approach for estimating the gap time hazard. Smoothing spline ANOVA decompositions are used to model the log gap time hazard as a joint function of gap time and covariates, and general frailty is introduced to account for

between-subject heterogeneity and within-subject correlation. We estimate the nonparametric gap time hazard function and parameters in the frailty distribution using a combination of the Newton-Raphson procedure, the stochastic approximation algorithm (SAA) and the Markov Chain Monte Carlo (MCMC) method. The convergence of the algorithm is guaranteed by decreasing the step size of parameter update and/or increasing the MCMC sample size along iterations. Model selection procedure is also developed to identify negligible components in a functional ANOVA decomposition of the log gap time hazard. We evaluate the proposed methods with simulation studies and illustrate its use through the analysis of bladder tumor data.

Insights On The Robust Variance Estimator Under Recurrent-Events Model

✦Hussein Al-Khalidi, Duke University, 2400 Pratt Street, Durham, NC 27710, *hussein.al-khalidi@duke.edu*

Key Words: Andersen-Gill model, Clinical trials, Recurrent events data, Robust standard error, Sample size, Sandwich estimator

Recurrent events are common in medical research for subjects who are followed for the duration of a study. For example, cardiovascular patients with an implantable cardioverter defibrillator (ICD) experience recurrent arrhythmic events which are terminated by shocks or antitachycardia pacing delivered by the device. In a published randomized clinical trial, a recurrent-event model was used to study the effect of a drug therapy in subjects with ICDs who were experiencing recurrent symptomatic arrhythmic events. Under this model, one expects the robust variance for the estimated treatment effect to diminish when the duration of the trial is extended, due to the additional events observed. However, as shown in this paper, that is not always the case. We investigate this phenomenon using large datasets from this arrhythmia trial and from a diabetes study, with some analytical results, as well as through simulations. Some insights are also provided on existing sample size formulae using our results.

Median Cost Analysis For Recurrent Event Data

✦Alexander McLain, NICHD, , *mclaina@mail.nih.gov*; Raji Sundaram, National Institute of Child Health & Human Development; Subhashis Ghoshal, North Carolina State University

Key Words: Cost analysis, Intensity function, Proportional rate model, Recurrent events

In biomedical studies subjects may experience event of interest repeatedly. Such recurrent event data have been much studied in the statistical literature with application to cancer tumors, small bowel motility, schizophrenia, serious AIDS infections and many others. In many of these examples the time to such episodic events are associated with medical costs in treating them. To understand the accumulated cost for treating such a subject an understanding of both the recurrent event process and the associated medical costs process is needed. Furthermore, the distributions of the number of recurrent events and the cost of each episode are commonly right skewed. For this reason the mean cost process may not be a representative estimate. Here we use nonparametric recurrent event survival analysis techniques and a Gamma cost model to estimate, and predict, aspects of the median cost distribution. We derive the asymptotic distribution of the mean and median cost, and apply our techniques to the SEER-Medicare data set. This is based on joint work with S. Ghoshal (NCSU) and R. Sundaram (NICHD/NIH).

Estimation of Medical Costs Associated with Recurrent Events in the Presence of a Terminating Event

✦Yu Ma, University of Michigan, 1023 Barton Drive, Apt 206, Ann Arbor, MI 48105, *rickma@umich.edu*; Douglas E. Schaubel, University of Michigan

Key Words: hierarchical modeling, terminating event, generalized estimating equation, proportional rate, martingale structure, proportional hazard

In the biostatistical literature, many methods have focused on estimating medical costs. However, few have been developed under a framework where subjects experience both recurrent events (e.g., hospitalizations) and a terminating event (e.g.,death); a frequently occurring data structure. We propose novel methods which contrast group-specific cumulative mean costs, contingent on recurrent event and survival experience. Our proposed methods utilize a form of hierarchical modeling: a proportional hazards model for the terminating event; a proportional rates model for the conditional recurrent event rate given survival; and a linear model for cost, given hospitalization. Mean cost, viewed as a process over time, is estimated by combining fitted values from each of the afore-listed models. Large sample properties are derived, while simulation studies are conducted to assess finite sample properties and to evaluate robustness under misspecified models. We apply the proposed methods to data obtained from the Kidney Epidemiology and Cost Center, which motivated our research.

A Marginal Approach For Multivariate Survival With Longitudinal Covariates

✦Yi-Kuan Tseng, Graduate Institute of Statistics, National Central University, Jhong-Li, Tauyuan, International 32054 Taiwan, *tsengyk@ncu.edu.tw*; Ya-Fang Yang, Graguate Institute of Statistics, National Central University

Key Words: Joint model, MCEM, mixed model, misspecification, Cancer vaccine

In clinical trials and other medical studies, it has become increasingly common to observe multiple event times of interest and longitudinal covariates simultaneously. In the literature, joint modeling approaches have been employed to analyze both survival and longitudinal processes and to investigate their association. Early attention has mostly been placed on developing adaptive and flexible longitudinal processes based on a prespecified univariate survival model, most commonly chosen as the Cox proportional model. We propose a marginal likelihood approach to handle multivariate survival time in joint model framework which implements the similar idea of marginal methods used in literature by ignoring the dependency among event times. The marginal likelihood could be easily incorporated various survival model in the likelihood function including two popular survival models, Cox and AFT models, or others such as extended hazard model. The maximization of the marginal likelihood is conducted through Monte Carlo EM and the standard error estimates are obtained via bootstrap method. The performance of the procedure is demonstrates through simulation study and case study.

Inference For Doubly Censored Data Using Marginal Likelihood

✦Zhiguo Li, Duke University, , *zhiguo.li@duke.edu*; Kouros Owzar, Duke University

Key Words: doubly censored data, marginal likelihood

In some applications, the variable of interest is time from a first event to a second event, while both times are interval censored. We propose fitting Cox proportional hazards model to this type of data using marginal likelihood, where the time to first event is integrated out in the empirical likelihood function of the time of interest. This greatly reduces the complexity of the likelihood function compared with the full semiparametric likelihood. The dependence of the time of interest on time to the first event (origin) is induced by including time to the first event as a covariate in the Cox model for the time of interest. Theory for the estimator is established and simulation is conducted to assess its performance. It's also applied to a real data set.

169 High Dimensional Graphical and Correlated Modeling

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

Monday, August 1, 10:30 a.m.–12:20 p.m.

Model Selection And Estimation In Matrix Normal Graphical Model

✦Jianxin Yin, University of Pennsylvania, 206 Blockley Hall, 423 Guardian drive, Philadelphia, PA 19104, *yinj@upenn.edu*; Hongzhe Li, University of Pennsylvania

Key Words: Guassian graphical model, l1 penalty function, matrix normal distribution, oracle property, pairwise markov property

Motivated by analysis of gene expression data measured over different tissues, we consider matrix-valued random variable and matrix-normal distribution and present a penalized estimation method for the corresponding concentration matrices. We show that the concentration matrices have a graphical interpretation for genes and tissues. We develop an efficient algorithm based on the graphical lasso to implement the penalized estimation. Asymptotic distribution of the estimates and the oracle properties are developed. Simulations and real examples demonstrate the competitive performance of the new methods.

Estimating Networks With Jumps

✦Mladen Kolar, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA USA, *mladenk@cs.cmu.edu*; Eric Poe Xing, Carnegie Mellon University

Key Words: Gaussian graphical models, network models, dynamic network models, high-dimensional inference, structural changes

We study the problem of estimating a temporally varying coefficient and varying structure graphical (VCVS) model underlying nonstationary time series data, such as social states of interacting individuals or microarray expression profiles of gene networks, as opposed to i.i.d. data from an invariant model widely considered in current literature of structural estimation. In particular, we consider the scenario in which the model evolves in a piece-wise constant fashion. We propose a procedure that minimizes the so-called TESLA loss (i.e., temporally smoothed L1 regularized regression), which allows jointly estimating the partition boundaries of the VCVS model and the coefficient of the sparse precision matrix on each block of the partition. A highly scalable proximal gradient method is proposed to solve the resultant convex optimization problem; and the conditions for sparsistent estimation and the convergence rate of both the partition boundaries and the network structure are established for the first time for such estimators.

Issues In Longitudinal Modeling Of Large Sparse Social Networks

✦Sudeshna Paul, Harvard Medical School, Department of Health Care Policy, 180 Longwood Avenue, Boston, MA 02115-5899, *paul@hcp.med.harvard.edu*; A. James O'Malley, Harvard Medical School

Key Words: dynamic networks, dyads, tie formation, tie dissolution, sparse, sample

Statistical modeling and analysis of dynamic networks has been a challenging area due to the complex structure and lack of manageable data. Recently we have proposed a log-linear model based on contingency tables to model transition probabilities of dyads (a pair of nodes) corresponding to tie formation and dissolution in a network. A serious problem arises in the case of applying this model to large networks with sparse connections. Frequently the 00|00 cell in the contingency table dominates the first row resulting in near zero 01|00, 10|00, 11|00 relative cell frequencies, and thus resulting in a model with inestimable parameters. To partially resolve this issue, we propose to fit the model on a subset of dyads in the complete network that include all "non null" dyads across time but only a random sample of "always" null dyads. To enable the uncertainty in the selection of always-null dyads, the process is iterated and summary estimates computed. Sensitivity analysis is performed to study the effects of the subset size and number of iterations on the parameter estimates.This would help in the extraction of some of the parameters of interest and reduce computation costs.

Practice-Related Changes In Neural Circuitry Supporting Eye Movements Investigated Via Wavelet-Based Clustering Analysis

✦Jinae Lee, University of Georgia, 259 Statistics Bldg., University of Georgia, Athens, GA 30602, *jinaelee@gmail.com*; Cheolwoo Park, University of Georgia; Benjamin Austin, University of Wisconsin; Kara Dyckman, University of Georgia; Qingyang Li, University of Georgia; Jennifer E. McDowell, University of Georgia; Nicole Lazar, Department of Statistics

Key Words: fMRI, temporal correlation, clustering, wavelet, saccade tasks, no trend test

In fMRI studies clustering methods are used to detect similarities in the activation time series among voxels. It is assumed that the temporal pattern of activation is organized in a spatially coherent fashion such that clustering will extract the main temporal patterns and partition the dataset by grouping similarly behaved functions together. We propose a clustering procedure built in the wavelet domain to take temporal correlation into account. We also construct a no trend test based on wavelets to significantly reduce the high dimension of the data prior

to clustering. In an actual clustering step, PCA K-means is applied and produces clustered maps that show the apparent structure of the activation of a brain. First, we evaluate the performance of our clustering method using the simulated data and compare it to other approaches. Second, we analyze the fMRI data acquired on two occasions while the 37 participants were engaged in saccade tasks (anti-saccade, pro-saccade, and fixation). We attempt to aggregate voxel time series into a small number of clusters and compare the clustered maps for the three practice groups and also between the two scan time points.

Robust Event-Related Fmri Designs Under A Nonlinear Model

✦Ming-Hung Kao, Arizona State University, , *mkao3@asu.edu*; Dibyen Majumdar, University of Illinois at Chicago; Abhyuday Mandal, University of Georgia; John Stufken, University of Georgia

Key Words: A-optimality, Genetic algorithms, Hemodynamic response function, Information matrix, Maximin efficient designs

Previous studies on event-related functional magnetic resonance imaging (ER-fMRI) experimental designs are primarily based on linear models, in which a known shape of the hemodynamic response function (HRF) is assumed. However, the HRF shape is usually uncertain at the design stage. To address this issue, we consider a nonlinear model to accommodate a wide spectrum of feasible HRF shapes, and propose an approach for obtaining maximin efficient designs. Our approach involves a reduction in the parameter space and an efficient search algorithm. The designs that we obtain are much more robust against mis-specified HRF shapes than designs widely used by researchers.

Model Selection and Goodness of Fit for Phylogenetic Comparative Methods

✦Dwueng-Chwuan Jhwueng, National Institute for Mathematical and Biological Synthesis, NIMBioS 1534 White Ave., Knoxville, TN 37996, *djhwueng@indiana.edu*

Key Words: phylogenetic comparative methods, Independent Contrasts, phylogenetic mixed model, Ornstein-Uhlenbeck (OU) process, spatial autoregressive model, Akaike information criterion

Phylogenetic comparative methods (PCMs) have been applied widely in analyzing data from related species. Many such methods have been proposed but their fit to data is rarely assessed. We assess the fit of several phylogenetic comparative methods to a large collection of real data sets gathered from the literature. We also compare the models using model selection criteria. Results show that Felsenstein's Independent Contrast and the independent, non-phylogenetic, models provide better fit for most real data. We then fit those methods to bivariate data sets and compare correlation estimates and confidence intervals from each model using simulations. We find that correlations from different models are often qualitatively similar so that actual correlations from real data seem to be robust to the PCM chosen for the analysis. Therefore, while there is no evidence that the parameter-richer models fit better, using them does not overly change the result in most bivariate analyses.

Mixing Times For A Class Of Markov Chains On Phylogenetic Tree Spaces

✦David Allen Spade, The Ohio State University Department of Statistics, 1958 Neil Avenue, Columbus, OH 43210, *spade.10@buckeyemail.osu.edu*; Radu Herbei, The Ohio State University; Laura Kubatko, The Ohio State University

Key Words: Markov Chains, Phylogenetic trees, Mixing Times

In the past decade, there has been a considerable amount of work done in studying the mixing time for Markov chains on phylogenetic trees. In the current work we extend existing results to Markov chains that are commonly used in algorithms for phylogenetic inference. Specifically, we study NNI, SPR and TBR moves on rooted and unrooted trees with n taxa. For such trees we estimate the total variation distance between the law of the current state and the stationary law in some simple cases and provide upper and lower bounds for this distance in the general case. We will also perform some simulations to investigate how well the bounds behave.

170 Solutions to Various Complexities in the Analysis of Health Data

Section on Health Policy Statistics, Scientific and Public Affairs Advisory Committee

Monday, August 1, 10:30 a.m.–12:20 p.m.

Bias In Variance Estimation Using Re-Sampling Of Longitudinal And Nested Administrative Health Data.

✦Bassam Dahman, Virginia Commonwealth University, Richmond, VA 23298 USA, *bdahman@msn.com*

Key Words: resampling, bias, longitudinal, nested, administrative data

Re-sampling methods are widely used in estimating the variance and standard errors of parameter estimates and predicted values, and in determining the statistical significance in hypothesis testing using administrative health data. In longitudinal and nested models, bias might be introduced to these estimates if the samples generated by the re-sampling routine differ in their distribution and longitudinal or nesting structure from the original data. This bias might be particularly large in the presence of missing data. In this paper we study and demonstrate the properties of different methods of re-sampling longitudinal and nested data. Using a simulation study based on longitudinal hospital level data and nested discharge data we compare between the different re-sampling routines, and evaluate the bias corrections required for each method.

The Association Of Hepatitis B Universal Supply Policy With Timing Of The First Dose Hepatitis B Shot, National Immunization Survey (Nis).

✦Zhen Zhao, Centers for Disease Control and Prevention, 1600 Clifton Road, MS E-62, Atlanta, GA 30333, *zaz0@cdc.gov*; Trudy Murphy, Centers for Disease Control and Prevention

Key Words: Hepatitis B, Universal Supply Policy, Cox model, Complex Survey

The association of state hepatitis B vaccine (HepB) universal supply policy with receipt of HepB within three days of birth has been assessed. However, those methods ignored information on timing of first HepB dose and did not adjust for covariates. This study applied Cox's Proportional Hazards model for complex survey data to evaluate the relationship between HepB universal supply policy and days from birth to receipt of the first dose of HepB in the first 30 days after birth. Data for infants born in 2006 from National Immunization Survey (NIS) were used. Proportional hazards assumptions were checked for each covariate. The results show the HepB universal supply policy was significantly associated with days from birth to first dose when controlling for provider participation in Vaccine for Children (VFC) program, number of providers, provider type, infant's race/ethnicity, and mother's education. Receipt of the first HepB dose occurred sooner in newborns living in a state with universal supply policy than in other states (hazard ratio 1.30, 95%CI 1.22-1.38) and sooner in newborns with providers participating in VFC than without such providers (hazard ratio 1.28, 95%CI 1.15-1.42).

Exact Logistic Regression For Clustered Data With With Varying Intra-Cluster Dispersion

✦Trent Lalonde, University of Northern Colorado, McKee Hall 520, Campus Box 124, Greeley, CO 80634, *trent.lalonde@unco.edu*; Jeffrey Wilson, Arizona State University

Key Words: Binary Response, Correlated Response, Logistic Regression, Exact Methods

Exact methods for binary responses are often necessary when asymptotic methods such as maximum likelihood fail for small sample sizes (Cox (1970)). For the situation of correlated data, logistic methods have been proposed for the asymptotic case (Connolly and Liang (1988)), and also for the exact case (Corcoran, Ryan, Senchaudhuri, Mehta, Patel, and Monenbergs(2001)). Exact logistic models for data showing multiple levels of clustering have been investigated by Troxler, Lalonde, and Wilson (2011). These methods have shown utility in modeling correlated binary responses, such as repeated observation of success for individuals. This paper extends a brief discussion of Troxler et al (2011) of exact logistic models that allow for varying magnitudes of correlation within clusters. This extension is important because it allows the data greater control in determining an appropriate model. The model is presented, and computational considerations are discussed for various cluster and sample sizes. A simulation study is presented to assess typical running times. Finally connections are made between varying intra-class correlation and overdispersion in logistic models.

Sample Size Determination For Hierarchical Longitudinal Designs

✦Joseph Warfield, Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Rd, Laurel, MD 20723, *joseph.warfield@jhuapl.edu*; Anindya Roy, University of Maryland Baltimore County

Key Words: Bernstein Polynomial, Mixed Effects, Power Analysis, Poisson Regression

The work in this paper considers the problem of sample size determination for mixed effects Poisson regression models for the analysis of clustered longitudinal data. The approach is to transform the count response data using Bernstein polynomials, which reduces the problem to a mixed-effects linear regression model. We derive sample size requirements (i.e., power characteristics) for a test of treatment-by-time interaction for designs with different randomization schemes.

Evaluation Of A Confidence Interval Approach For Absolute Inter-Rater Reliability In A Crossed Three-Way Random Effects Model

✦Joseph C Cappelleri, Pfizer Inc, 50 Pequot Ave, New London, CT 06320 United States, *joseph.c.cappelleri@pfizer.com*; Naitee Ting, Boehringer Ingelheim Pharmecuticals

Key Words: inter-rater reliability, intraclass correlation coefficient, confidence interval;, analysis of variance, Monte Carlo simulation;, generalizability theory.

We specify a three-factor random-effects model from an inter-rater reliability study, where the effects of subjects, raters, and items are random. The reliability measure is an intraclass correlation coefficient that measures the absolute agreement of a single measurement from one rater on an item to another rater on the same item. Our objective is to evaluate and illustrate an approximate confidence interval around this intraclass correlation coefficient based on Satterthwaite's approximation (Wong and McGraw. Educational and Psychological Measurement 1999; 59:270-288). In doing so, we performed Monte Carlo stimulations and provided two examples. Overall, coverage of 95% one-sided lower bounds and upper bounds, along with 90% confidence intervals, maintained their true coverage or were slightly conservative. This investigation is the first to validate the methodology.

A Sensitivity Analysis Addressing Missing Data Within The Young Lives Longitudinal Study

✦Mark Griffin, School of Population Health, University of Queensland, Herston, Brisbane, International 4006 Australia, *m.griffin@uq.edu.au*; Rosa Alati, School of Population Health, University of Queensland; Abdul Mamun, School of Population Health, University of Queensland; Rob Ware, School of Population Health, University of Queensland

Key Words: missing data, longitudinal, sensitivity, poverty, community aid

In this presentation I will compare the results of two methods for handling missing data within an analysis of the Young Lives longitudinal study. These two likelihood-based methods are a Missing At Random method (Lipsitz, Biometrika 1996) and a Missing Not At Random method (Stubbendick, Biometrics 2003). These methods have been applied to a setting where the covariates as well as the study outcome contain missing values. Through comparing the results from these methods we will explore the sensitivity of the study results to assumptions about the mechanism of missingness. This comparison will be conducted within an analysis of the Young Lives longitudinal study, an ongoing study of child poverty in four countries (Ethiopia, India, Peru, and Vietnam). 2000 children born in 2001-2 and 1000 children born in 1994-5 were selected in each country, making a total

of 12,000 children that will be followed for a total of 15 years. Children are the largest age group affected by poverty, an end to poverty has to start with them.

A Comparison Of Analyses For Two Group Small Samples With A Large Number Of Measures

✦Margo Sidell, Tulane University, 6058 Camp St, New orleans, LA 70118 USA, *msidell@tulane.edu*; Leann Myers, Tulane University

Key Words: small sample size, multivariate analysis

In many fields, data are collected that have a greater number of continuous variables (p) than group sample size (n). The classical two sample method for comparing multivariate mean vectors, Hotelling's T2 test, is undefined when p > 2n - 2. This study examined four alternative two sample tests that are not restricted by the number of variables: a component wise statistic (Wu, Genton, & Stefanski, 2006), an extension of Hotelling's T2 test (Schott, 2007), an ANOVA type nonparametric test (Bathke, Harrar, & Madden, 2008), and a new two sample U-score test based on a single sample U-score statistic (Wittkowski et al., 2008). Monte Carlo simulations were run for each of the four tests. Multivariate normal data were generated with low, moderate, and high correlation where the number of variables was equal to or exceeded the group sample size. Each test was used for each dataset. This study investigated which of these four methods was most appropriate in terms of type I error and power under various conditions.

171 Empirical Likelihood and Robust Methods ●

Section on Nonparametric Statistics

Monday, August 1, 10:30 a.m.–12:20 p.m.

On Maximum Empirical Likelihood Estimation And Related Topics

✦Hanxiang Peng, Indiana University Purdue University Indianapolis, IN 46202 USA, *hpeng@math.iupui.edu*; Anton Schick, Binghamton University

Key Words: irregular constraints, nuisance parameters, empirical likelihood ratio tests, local asymptotic normality condition, maximum empirical likelihood estimators, efficiency

This article studies maximum empirical likelihood estimation in the case of constraint functions that may be discontinuous and/or depend on additional parameters. The later is the case in applications to semiparametric models where the constraint functions may depend on the nuisance parameter. Our results are thus formulated for empirical likelihoods based on estimated constraint functions that may also be irregular. The key to our analysis is a uniform local asymptotic normality condition for the local empirical likelihood ratio. This condition holds under mild assumptions on the estimated constraint functions and allows for a study of maximum empirical likelihood estimation and empirical likelihood ratio testing similar to that for parametric models with the uniform local asymptotic normality condition. Applications of our results are discussed to inference problems about quantiles under possibly additional information on the underlying distribution, to residual-based inference about quantiles, and to partial adaption.

Empirical Likelihood Approximation to Neyman-Pearson Tests for Sample Distributions

✦Albert Vexler, The New York State University at Buffalo, NY 14214 USA, *avexler@buffalo.edu*; Gregory Gurevich, Department of Industrial Engineering and Management, SCE- Shamoon College of Engineering

Key Words: Empirical likelihood, Entropy, Likelihood ratio, Two-sample nonparametric tests

In this paper we develop two-sample empirical likelihood approximations to parametric likelihood ratios, resulting in an efficient test based on samples entropy. The proposed and examined distribution-free two-sample test is shown to be very competitive with well known nonparametric tests. For example, the new test has high and stable power detecting a nonconstant shift in the two-sample problem, when Wilcoxon's test may break down completely. This is partly due to the inherent structure developed within Neyman-Pearson type lemmas. The outputs of a broad Monte Carlo analysis support our theoretical results and indicate that the proposed test compares favorably with the standard procedures, for a wide range of null and alternative distributions.

Empirical Likelihood Confidence Intervals For Roc Curves With Missing Data

✦Yichuan Zhao, Georgia State University, Department of Mathematics and Statistics, Atlanta, GA 30303 USA, *dz2007@gmail.com*; Yueheng An, Georgia State University

Key Words: Confidence interval, Missing data, ROC curve, Smoothed empirical likelihood

The receiver operating characteristic (ROC) curve is widely utilized to evaluate the diagnostic performance of a test, in other words, the accuracy of a test to discriminate normal cases from diseased case. In biomedical studies, we often meet with missing data. In this situation, the regular inference procedures cannot be applied directly. In this paper, a random hot deck imputation is used to obtain a "complete sample". After that, empirical likelihood (EL) confidence intervals are constructed for the ROC curves. The empirical log-likelihood ratio statistics is derived whose asymptotic distribution is a weighted chi-square distribution. The results of simulation study show that the EL confidence intervals perform well in terms of the coverage probability and average length for various sample sizes and respondent rates.

Improved Polynomially-Adjusted Density Estimates

✦Serge B. Provost, The University of Western Ontario, Dept. of Stat. & Act. Sciences, WSC Room 262, London, ON N6A5B7 Canada, *provost@stats.uwo.ca*; Min Jiang, Statistics Canada

Key Words: Density estimation, Orthogonal polynomials, Degree selection, Multivariate distributions

Density estimates that are expressible as the product of a base density function and a linear combination of orthogonal polynomials are being considered. More specifically, two criteria are proposed for determining the number of terms to be included in the polynomial adjustment component and guidelines are suggested for the selection of a suitable base density function. A simulation study reveals that these stopping rules produce density estimates that are generally more accurate than kernel density estimates or those resulting from the application of the Kronmal-Tarter criterion. Additionally, it is explained that the same approach can be utilized to obtain multivariate density estimates. The proposed orthogonal polynomial density estimation methodology is applied to several univariate and bivariate data sets.

Robustness of Random Forests for Regression

✦Denis Larocque, HEC Montreal, , *denis.larocque@hec.ca*; Marie-Helene Roy, HEC Montreal

Key Words: Random forest, Robustness, Median, Ranks

In this paper, we empirically investigate the robustness of random forests for regression problems. We also investigate the performance of five variations of the original random forest method, all aimed at improving robustness. All the proposed variations can be easily implemented using the R package randomForest. The competing methods are compared via a simulation study and ten real data sets obtained from the UCI Machine Learning Repository. Our results show that the median-based random forests offer good and stable performances for the simulated and real data sets considered and, as such, should be considered as serious alternatives to the original random forest method.

Invariant Estimation Of Location And Scale Parameters In Generalized Skew-Symmetric Distributions

✦Nelis Potgieter, Texas A&M University, , *nelis@stat.tamu.edu*; Marc G. Genton, Texas A&M University

Key Words: skew-normal, skew-t, characteristic function, root selection

In generalized skew-symmetric models, it is of interest to estimate the location and scale parameters without assuming a specific parametric form for the skewing function. In considering this problem, it has been established that the there are typically multiple solutions for the parameters of interest. The problem of selecting the "correct" root has only been solved satisfactorily in a few cases. We present a new method for estimting the parameters, namely using a distance function based on the real parts of the empirical and true characteristic functions. We proceed to show that the multiple roots that occur are a result of an identifiability issue with the fully generalized skew-symmetric model. Making an additional assumption regarding the behaviour of the skewing function in a region around the origin, this can be overcome. Theoretical and Monte Carlo standard errors for the parameter estimates are used to compare the new method to existing methods.

A New Method for Generating Families of Continuous Distributions

✦Ayman Alzaatreh, Central Michigan University, 1298 Granger St., Mount Pleasant, MI 48858, *alzaa1ay@cmcih.edu*; Carl Lee, Central Michigan University; Felix Famoye, Central Michigan University

Key Words: hazard function, generalized distribution, beta-family

A new method is proposed to generate a family of continuous distributions. A random variable X, "the transformer", is used to transform another random variable T, "the transformed". The resulting family, the T-X family of distributions, has a connection with the hazard functions and each generated distribution is considered as a weighted hazard function of the random variable X. Many new distributions, which are members of the family, are presented. Several known continuous distributions are found to be special cases of the new distributions.

A Dominated Rejection Algorithm For Generating Random Variates

✦Timothy Hall, PQI Consulting, P. O. Box 425616, Cambridge, MA 02142-0012, *info@pqic.com*

Key Words: generating random variates, uniform variate, embedded systems applications

This paper presents a practical modified version of the von Neumann Dominated Rejection Method for generating univariate random variates using a distibution density function, a uniform variate over a finite interval, and an independent uniform variate over the unit interval. Several example generated variates from the normal distribution family are included for demonstration purposes. The algorithm segments used to implement this version are presented as MMIX code for use in creating embedded systems applications.

A Family Of Asymmetric Distributions On The Circle

✦SHONGKOUR ROY, JAHANGIRNAGAR UNIVERSITY, M.S STUDENT, DEPARTMENT OF STATISTICS, SAVAR, DHAKA, International 1342 BANGLADESH, *sankar1604@gmail.com*; Mian Arif Shams Adnan, JAHANGIRNAGAR UNIVERSITY

Key Words: Angular data, Burr distribution, Extreme value distribution, Lerch's function

Application of directional data analysis is getting more and more important in many scientific disciplines like Biophysics, Astrophysics, Medicine, Biology, Geology etc. But the available wrapped symmetric distributions are not sufficient enough to represent many natural phenomenons of the circular data. All previously derived distributions do not include Wrapped Extreme Value, Wrapped Maxwell, Wrapped Rayleigh, Wrapped Burr, etc. The appropriate probabilistic models of all the above distributions have been derived. Estimation of unknown parameters along with the basic belongings of these distributions is also provided. The shapes of the aforementioned distributions are also investigated through the graphical representation and tabular exhibition for various specifications of the concern parameters.

172 Distributions and Variate Generation

Section on Statistical Computing, Section on Statistical Graphics
Monday, August 1, 10:30 a.m.–12:20 p.m.

High Dimensional Generation Of Bernoulli Random Vectors

✦Reza Modarres, George Washington University, Department of Statistics, 2040 Pennsylvania Ave, Washington DC, DC 20052, *reza@gwu.edu*

Key Words: Multivariate, Bernoulli, Mixture Model, Latent Variable, Random Vectors

We explore different modeling strategies to generate high dimensional Bernoulli vectors, discuss the multivariate Bernoulli (MB) distribution, probe its properties and examine three models for generating random vectors. A latent multivariate normal model whose bivariate distributions are approximated with Plackett distributions with univariate normal distributions is presented. A conditional mean model is examined where the conditional probability of success depends on previous history of successes. A mixture of Beta distributions is also presented that expresses the probability of the MB vector as a product of correlated binary random variables.

A Class Of Triple-Mixture Distributions

✦HUMAYUN KISER, JAHANGIRNAGAR UNIVERSITY, DEPARTMENT OF STATISTICS, SAVAR, DHAKA, International 1342 BANGLADESH, *humayun_kiser@yahoo.com*; Mian Arif Shams Adnan, JAHANGIRNAGAR UNIVERSITY

Key Words: Mixtured distribution, Mixing distribution

We have derived Triple-mixture distributions of most of the discrete and continuous distributions each of which Triple-mixture distributions is a continuous mixture of three same distributions. Estimation of unknown parameters along with some characteristics of these distributions is also investigated.

Generalized Variable Approach For Weibull Distributions

✦Yin Lin, Northern Arizona University, P.O.Box 5717, Department of Mathematics & Statistics, Flagstaff, AZ 86011, *yinlin7272@gmail.com*

Key Words: Weibull Distributions, Type I censored, generalized variable approach, confidence interval

This talk considers inferential procedures for Weibull distributions based on type I censored samples. The methods are based on the classical pivotal quantities and the concept of generalized variable approach. In some cases the methods are exact when samples are type II censored, and they can be used as approximates for type I censored samples. The proposed methods will be illustrated for constructing confidence intervals for the difference between or the ratio of two Weibull means. The methods are are illustrated using some practical examples.

A Generalization Of The Family Of Exponential And Beta Distributions

✦Mian Arif Shams Adnan, JAHANGIRNAGAR UNIVERSITY, ASSISTANT PROFESSOR, DEPARTMENT OF STATISTICS, SAVAR, DHAKA, 1342 BANGLADESH, *julias284@yahoo.com*; HUMAYUN KISER, JAHANGIRNAGAR UNIVERSITY

Key Words: Maximum Likelihood Function, Moment, Shape parameter

The generalized forms of a set of distributions like Exponential, Gamma, Chi-square, t, F, Beta first kind, Beta second kind, etc have been suggested. The basic belongings and the maximum likelihood estimators of the unknown parameters of all the generalized forms have been found.

173 Section on Survey Research Methods-- Small area estimation in theory and practice

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

Monday, August 1, 10:30 a.m.–12:20 p.m.

Estimation Of Poverty At The School District Level Using Reweighting Method

✦Sam Hawala, U.S. Census Bureau, 4600 Silver Hill Road, Room 6H124F, Suitland, MD 20233, *sam.hawala@census.gov*; Partha Lahiri, University of Maryland at College Park

Key Words: Structure Preserving Estimation, Synthetic Estimation, Average mean Square Error

In this paper, we adapt a reweighting method of Schirm and Zaslavsky (1999) to estimate poverty at the school district level using the American Community Survey data in conjunction with administrative records. The proposed method distributes the weight associated with a person among other school districts based on the similarity of the person with those in the other school districts. This process of borrowing strength is especially important for the poor who may only have a few sample observations in a school district. We compare the results to the Bureau's current SAIPE approach for school district estimation. In particular, we compare the average design-based mean squared error estimates of different estimates.

A Bayesian Zero-One Inflated Beta Model For Estimating Poverty In U.S. Counties

✦Jerzy Wieczorek, U.S. Census Bureau, 4600 Silver Hill Road, Room 6H124C, Suitland, MD 20233, *jerzy.wieczorek@census.gov*; Sam Hawala, U.S. Census Bureau

Key Words: Small Area Estimates, SAIPE, MCMC, Beta Regression, Hierarchical Model

We propose and evaluate a Bayesian beta regression model for U.S. county poverty rates. Such a rate model could be an improvement to the Census Bureau's current small-area poverty approach of linearly modeling the logarithm of poverty levels. For small areas, some of which may have estimates of no poverty or all poverty, a zero-one inflated rate model can usefully account for estimated rates of 0 or 1. Using Bayesian computation techniques, we estimate the parameters of a zero-one inflated beta regression model. We compare the results to the Census Bureau's current small-area model for county poverty estimation.

Application Of Small Area Estimation For Annual Survey Of Employment And Payroll

✦Bac Tran, U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20746, *bac.tran@census.gov*

Key Words: Decision-based Estimation, Modified Direct Estimator, Synthetic Estimation, Composite Estimation

Annual Survey of Employment and Payroll estimates the number of federal, state, and local government employees and their gross payrolls. In the past two years, we developed the decision-based method to estimate the survey total. In this paper, we discuss some small area challenges when we estimate the survey total at the functional level of government units such as airport, public welfare, hospitals, etc. First, we introduce the synthetic estimation and modified direct estimators. Then, we modified the composite estimation as a weighted average between modified direct estimation and synthetic estimation. Finally, we evaluate these methods by using the 2007 Census of Governments: Employment Component.

An Empirical Best Linear Unbiased Prediction Approach To Small Area Estimation Of Crop Parameters

✦Michael E. Bellow, USDA-NASS, 3251 Old Lee Highway, Rm. 305, Fairfax, VA 22031, *mbellow@nass.usda.gov*; Partha Lahiri, University of Maryland at College Park

Key Words: small area estimation, components of variance, predictor variables, AWiFS

Accurate county (small area) level estimation of crop and livestock items is an important priority for the USDA's National Agricultural Statistics Service (NASS). We consider an empirical best linear unbiased prediction (EBLUP) method for combining multiple data sources to estimate crop harvested area, yield and production at the county level. The method employs a unit level linear mixed model, with the variance components estimated using a new technique which (unlike the standard maximum likelihood and restricted maximum likelihood methods) ensures strictly positive consistent estimation of the model variance. In order to produce uncertainty measures of the proposed estimator and associated confidence intervals, a parametric bootstrap method that incorporates all sources of uncertainty is proposed. Related model diagnostics are also described, and results of a study evaluating the EBLUP method for corn and soybeans in seven midwestern states are discussed.

A "Virtual Population" Approach To Small Area Estimation

✦Michael P. Battaglia, Abt Associates, 55 Wheeler Street, Cambridge, MA 02138, *mike_battaglia@abtassoc.com*; Martin R. Frankel, Baruch College, CUNY; Lina S Balluz, Centers for Disease Control and Prevention

Key Words: Small area estimation, American Community Survey, Iterative Probability Adjustment (IPA), Health risk factors, County estimates

The BRFSS provides state estimates of health risk behaviors. Public health policy often requires health estimates at a level smaller that entire states. The BRFSS has embarked on a program to produce health measures at the individual county level. For the SAE system the micro level data from the American Community Survey serves as a "virtual population" for each state. For each risk factor a logistic regression model is developed based on the most recent BRFSS data for the state. Using this model, probabilities are assigned to individuals in the virtual population. Next direct estimates are developed for the entire state as well as by demographic characteristics. For example estimates of smoking would be developed for all persons in the state by Age, Gender, Race, and Education. Finally, using an iterative probability adjustment algorithm, the probabilities assigned to individuals in the virtual population are calibrated so that they sum to the survey based estimates in total and by demographic group. Upon convergence of the IPA, a simple cross-tabulation system may be used to obtain any specific estimate at the county level or any aggregation of counties.

Additive Random Coefficient (Arc) Models For Robust Small Area Estimation (Sae)

✦Ralph E. Folsom, RTI International, 3040 Cornwallis Road, Durham, NC 27709 USA, *ref@rti.org*; Akhil K. Vaish, RTI International; Avinash C Singh, NORC at the University of Chicago

Key Words: small area estimation, generalized design effects, nonignorable survey sampling design, general liner mixed model, additive random coefficient, survey weighted estimating equations

Unit or person-level ARC models with linear, logistic, and log-linear marginal mean functions are developed for SAE. ARC models take the form of a first order Taylor series approximation to the associated general linear mixed model. The area-level random coefficient vectors specify effects for demographic groups. Protection against nonignorable sample designs is provided by a hybrid solution that combines the marginal [probability(P) sampling plus ARC model(?)] distribution of the fixed regression coefficients with the MCMC simulated Bayes posterior distributions for the small area specific random coefficient vectors. Survey weighted estimating equations are employed in the solution for the fixed and random coefficients along with sample design consistent covariance matrix estimators. A generalized design effect matrix is used to smooth the area-level covariance matrices for the random coefficients. Nationally or regionally benchmarked solutions are specified. A simulation study for the logistic ARC model contrasts the new method's performance with some solutions currently in use that discount the effect of nonignorable samples on the mean squared errors of small area estimates.

Quasi-Blups For Reducing Over-Shrinkage In Small Area Estimation

✦Avinash C Singh, NORC at the University of Chicago, 55 East Monroe Street, 30th floor, Chicago, IL 60603, *singh-avi@norc.org*; Pin Yuan, Human Resources and Skills Development Canada

Key Words: Alternatives to EBLUP, Over-shrinkage, Zero or Negative Estimated Variance Component

In small area modeling, estimation of second order parameters (such as variance components or correlation coefficients in time series or spatial models) is often challenging because the estimates may turn out to be inadmissible or unreasonable. Estimated variance components may become very close to zero or negative (in which case it is truncated to zero or modified to get a positive estimate) due to model misspecifications

or due to large sampling errors. The resulting SAEs tend to exhibit over-shrinkage to synthetic estimates and may be far from the direct estimator. We propose quasi-BLUP estimation in which suitably pre-specified values are used for variance components for computing SAE but the MSE estimates are adjusted for using working values which may not be consistent. Empirical results in the context of Canadian LFS show that such estimates have desirable properties.

174 Section on Survey Research Methods - Evaluating Sample Designs and Redesigns

Section on Survey Research Methods, Section on Government Statistics

Monday, August 1, 10:30 a.m.–12:20 p.m.

Sampling Design for the 2010-2012 National Hospital Ambulatory Medical Care Survey

✦Iris Shimizu, National Center for Health Statistics, 8513 Montpelier Drive, Laurel, MD 20708, *IShimizu@cdc.gov*

Key Words: survey design, sampling design, health care survey

The National Center for Health Statistics (NCHS) conducts the National Hospital Ambulatory Medical Care Survey (NHAMCS) to measure utilization of ambulatory medical care service provided in non-Federal, non-institutional general and short-stay hospitals located in the 50 states and the District of Columbia. From its beginning in 1991-92, NHAMCS has collected data about sample visits made to hospital emergency and outpatient departments. In 2009, the survey also began collecting data about visits made to hospital ambulatory surgery centers (ASCs). In 2010, the survey universe was further expanded to include visits made to freestanding (not-hospital based) ambulatory surgery centers. This paper will discuss the sampling design for the current survey and differences between the survey's current and the original designs.

Innovative Northern Design Improvements In The Canadian Labour Force Survey

✦Edward J Chen, Statistics Canada, 18th Floor, RHC Bldg, Tunney's Pasture, Ottawa, ON K1A 076 Canada, *edward.chen@statcan.gc.ca*

Key Words: sample design, low population, data quality

by Scott Meyer and Edward J. Chen, Statistics Canada The Canadian Labour Force Survey (LFS) first started sampling in the Yukon Territory in 1991 with some earlier attempts made on a trial basis. Over time, the sample design has expanded to include Northwest Territories and Nunavut in conjunction with other household surveys conducted by Statistics Canada. The low population and vast land areas in the territories require alternate approaches from what has been implemented for the provinces. Of particular concern is the rate of use of the available sample. During the past two decades, there have been a number of sample design updates and coverage improvements in the North to improve the data quality of the survey estimates. This presentation describes the basic elements of the Northern Design and gives a summary of the design updates and improvements, focusing on the most recent adjustments which began to be phased-in starting in January, 2011.

Early Childhood Longitudinal Study: Kindergarten Class Of 2010-2011 - Sample Design Issues

✦Thanh LÍ, Westat, 1600 Research Blvd, Rockville, MD 20850, *thanhle@westat.com*; Greg Norman, Westat; Karen Tourangeau, Westat; J. Michael Brick, Westat Inc.; Gail Mulligan , National Center for Education Statistics

Key Words: Longitudinal survey, multi-stage sampling, oversampling, sample coverage, sample attrition

The Early Childhood Longitudinal Study: Kindergarten Class of 2010-2011 (ECLS-K) is the second longitudinal study of kindergartners sponsored by the National Center for Education Statistics. As with the 1998-99 cohort study, it will provide national data on children's characteristics as they progress from kindergarten through the fifth grade, and information on key analytical issues such as school readiness and transition from kindergarten to subsequent grades. Unlike the 1998-99 study where data were collected every other year after first grade, the 2010-2011 study aims to study kindergartners at every grade after kindergarten. In this paper, we discuss the sample design, describe school sampling frames, present procedures adopted to improve the school coverage, and discuss deviations from the 1998-99 sample design. The difficulty of implementing the sample - the recruitment of schools and parents - will also be presented in comparison with the 1998-99 study.

Sample Design Changes For The Nhanes 2011-2014 Annual Surveys

✦Hongsheng Hao, Westat, 1600 Research Blvd, Rockville, MD 20850, *hongshenghao@westat.com*; Leyla K. Mohadjer, Westat; Lin Li, Westat; Wendy Van de Kerckhove, Westat; Sylvia Dohrmann, Westat; Lester R. Curtin, National Center for Health Statistics

Key Words: Sample design, stratification, oversampling, household surveys

The National Health and Nutrition Examination Survey (NHANES) is an ongoing annual survey with PSUs selected for four years at a time. This process minimizes overlap in annual samples and allows for adjusting the design overtime to address changes in the national health concerns. A major design change for the 2011-2014 sample reflected the decision to oversample the Asian population to produce adequate sample sizes for given subdomains. In addition, the stratification scheme for sampling the PSUs was re-designed to ensure PSUs comprising annual and multi-year samples are distributed evenly in terms of health level, geography, urban-rural distribution, and population characteristics. The re-designed PSU stratification scheme was based on grouping the states by health-related measures. Cluster and factor analysis were used to form state groups using a number of state-level health indicators. Within each state group, NSR PSUs were further subdivided into substrata by geography, urban-rural distribution of population, race/ethnicity density, and poverty level. Various options of strata formation methods were performed and compared in the sample design to reach the design goal.

Predicting Violent Crime Rates For The 2010 Redesign Of The National Crime Victimization Survey (Ncvs)

✦Robert E. Fay , Westat, 1600 Research Blvd, Rockville, MD 20850, *bobfay@westat.com*; Jianzhu Li, Westat

Key Words: Sample design, UCR

The National Crime Victimization Survey (NCVS) is a major crime survey for the United States. The survey collects data on several types of crimes, including the broad categories of violent crime and property crime. The 2010 redesign of the NCVS can potentially improve the efficiency of the survey if the level of crime can be predicted well by external data. Previously, we reported initial success in predicting the level of crime at the county level based on the Uniform Crime Reporting (UCR) System. A more fine-grained analysis shows, however, far greater success for property crime than for violent crime. This paper extends the previous results to examine the underlying associations more thoroughly. We find that the largest single component of violent crime in the UCR, aggravated assault, fails to add to the predictive accuracy of a regression equation including the other violent crime components of the UCR, rape and robbery. We believe this finding has implications for interpretation of the UCR. We extend the analysis to include demographic characteristics from the census and the ACS, and we examine the ability of tract-level characteristics to predict individual victimizations.

Update On The Evaluation Of Sample Design Issues In The National Compensation Survey

✦Gwyn R Ferguson, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Room 3160, Washington, DC 20212, *ferguson.gwyn@bls.gov*; Chester Ponikowski, Bureau of Labor Statistics; Joan Coleman, U.S. Bureau of Labor Statistics

Key Words: survey design, sample allocation, dependent sampling, respondent burden, sample rotation

The National Compensation Survey is conducted by the Bureau of Labor Statistics to compute measures of the pay and benefits for America's workers. The current survey design uses a three-stage sample design to select samples of areas, establishments, and jobs for which wage and benefit data are collected periodically over a five-year rotation. In recent years, several potential changes to this design have been explored to increase survey efficiency, adjust to budget changes, reduce respondent burden, and reduce design complexity. Design areas that have been studied include sample rotation, allocation, sample frame preparation, establishment selection, and sample initiation scheduling. This paper will update the discussion of these issues, describe the alternative approaches that have been explored, present results from the recent design research, and present the recommended changes to the general survey design. The work in this paper updates and significantly expands upon the work presented in 2010 JSM Paper "Evaluating Sample Design Issues in the National Compensation Survey".

Alternative Probability Proportionate To Size Sampling Methods For The Ipp Sample Design

✦James Himelein, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Washington, DC 20212, *Himelein.Jim@bls.gov*

Key Words: Probability proportionate to size sampling, systematic vs. alternative procedures, Sunter method, Sampford Method, Yates-Grundy variance estimator

Probability proportionate to size without replacement (PPS WOR) sampling methods are customarily applied in sample designs for establishment surveys that have skewed populations with readily available auxiliary data for unit measures. More often than otherwise, systematic PPS procedures are used because such methods are easy to implement as well as being relatively efficient. However, there are alternative PPS sampling methods that are as efficient and relatively easy to implement that also afford the use of the Yates-Grundy variance estimator. This is demonstrated using the sample design from the International Price Program (IPP) at BLS.

175 Value-Added, IRT Models, and Education Outcomes

Value-Added, IRT Models, and Education Outcomes

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

Monday, August 1, 10:30 a.m.–12:20 p.m.

Exploring Missing Data In Value-Added Assessment Models In Education

✦Andrew Karl, Arizona State University, School of Math and Stats, Arizona State University, Tempe, AZ 85287-1804, *akarl@asu.edu*; Yan Yang, Arizona State University; Sharon Lohr, Arizona State University

Key Words: generalized linear mixed model, maximum likelihood, non-response model, random effects

The Federal Race to the Top Program encourages state education departments to develop longitudinal student records, following students even as they move between schools and districts. These records (e.g. test scores) are used in Value-Added Assessment (VAA) models to estimate the value added by individual teachers to student learning. Missing data occurs frequently in such longitudinal systems: students move out-of-state, are absent on the day of an examination, or drop-out of school. The VAA model's estimates for effects of teachers on students may be biased if the missing data is not missing at random. We present a new model for missing data that evaluates the potential effects of teachers on missing data and explore the properties and implications of the model.

Modeling Score-Based Student Achievement Data With Many Ceiling Values

✦Yan Yang, Arizona State University, School of Math and Stats, Arizona State University, Tempe, AZ 85287-1804, *yy@math.asu.edu*

Key Words: Ceiling effects, Censoring, Mixed effects, Value-added modeling

The No Child Left Behind Act of 2001 and the federal Race to the Top grant program initiated in 2009 both rely on standards-based state assessments to measure student achievement and school performance. In a standards-based test some students may attain the maximum possible score on an academic subject. This feature of the data not only invalidates the normality assumption widely adopted for modeling score-based student achievement data, but also manifests lost information on student learning. We develop a multilevel Tobit model that explicitly accounts for score ceilings in value-added assessment of school and teacher effects. Simulation and analysis of data from a standardized state assessment will be presented to demonstrate the practical utility of the proposed methods.

Teachers' Attitudes Toward Job Conditions: An Index Created By An Unfolding Irt Model

✦Weiwei Cui, National Institutes of Statistical Sciences, 1990 K St NW Suite 500, Washington, DC 20006 USA, *wcui@niss.org*

Key Words: Item Response Theory, Survey, Job Satisfaction

In the United States, schools chronically experience a shortage of qualified teachers. Ingersoll (2001) argues that this teacher shortage does not come from insufficient numbers of qualified teachers in the general population but rather from higher employee turnover in education than in other professions. Attitudes toward job conditions, such as job satisfaction, are found to be positively related to job performance (Judge, Bono, Thoresen, & Patton, 2001) and negatively associated with employee turnover (Harrison, Newman, & Roth 2006). This study applied the Generalized Graded Unfolding Model, a unidimensional parametric Item Response Theory (IRT) based unfolding model, to the teacher data collected by the Schools and Staff Survey (SASS), sponsored by National Center for Education Statistics. An index of job satisfaction was developed for each teacher in the dataset. This index provides an estimate of teachers' job satisfaction, and can be used to predict teacher turnover.

Assessing Goodness Of Fit Of Item Response Theory Models Using Generalized Residuals

✦Sandip Sinharay, Educational Testing Service, Rosedale Road, MS 12-T, Princeton, NJ 08541, *ssinharay@ets.org*; Shelby J. Haberman, Educational Testing Service

Key Words: model fit, psychometrics, educational statistics, item response theory

Item response theory (IRT) models (e.g., Lord, 1980), which are latent structure models in which manifest variables are polytomous and the latent variable/vector is polytomous or continuous, are often applied to scores on questions/items on standardized achievement or aptitude tests (Junker, 1993). Despite their long history, it is often not obvious how to rigorously assess the discrepancies between such models and observed data (Hambleton & Han, 2005). This situation exists despite the fact that Standard 3.9 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999) demands evidence of model fit when an IRT model is used to make inferences from a test data set. Generalized residuals are a tool employed in the analysis of contingency tables to examine goodness of fit. The essential feature of these residuals is that a linear combination of observed frequencies is compared to its estimated expected value under a proposed model. We apply these residuals to IRT models. Their use is illustrated with data from operational testing programs.

Modeling Outcomes for Elementary Science Education: The Science Writing Heuristic Project

✦Mack Shelley, Iowa State University, 1413 Snedecor Hall, Department of Statistics, Ames, IA 50011-1210, *mshelley@iastate.edu*; Christopher Gonwa-Reeves, Iowa State University; Joan Baenziger, Iowa State University; Brian Hand, University of Iowa; William Therrien, University of Iowa

Key Words: Multilevel modeling, Structural equation models, Science eduation, Science Writing Heuristic, Iowa Tests of Basic Skills, Cornell Critical Thinking test

An experimental-design study of the effects of the Science Writing Heuristic approach to providing elementary science instruction on student science content knowledge and critical thinking skills was implemented in 48 elementary school buildings in Iowa, with cluster random assignment of buildings to treatment and control groups based on percentage of students eligible for free and reduced lunch, enrollment in third through fifth grades, and private vs. public status. Confirmatory factor analysis of Level-1 (student) and Level-2 (building) characteristics for enhancing child outcomes undertaken using Mplus software shows statistically significant ($p < .05$) direct effects on students' Iowa Tests of Basic Skills results in mathematics, science, and reading comprehension from race, sex, free and reduced lunch eligibility, English language learner status, gifted and talented status, special education status, mathematics courses taken, and language instruction. Implications are discussed for advanced multivariate statistical model estimation, employing structural equation and hierarchical linear model methods.

Dropout Factories And College Enrollment: How School-Level Rates Of On-Time Grade Promotion Affect Matriculation

✦Thomas Christopher West, University of Delaware, 1015 Christina Mill Drive, Newark, DE 19711, *tcwest@udel.edu*

Key Words: administrative data, applied, secondary education, post-secondary education, education, grade retention

By utilizing Promoting Power as a school-level contextual variable, this study examines the relationship between school rates of on-time promotion and student matriculation. Specifically, this study uses a two-level hierarchical general linear model (HGLM) analytic design to identify significant student- and school-based factors that predict college enrollment behavior among U.S. Department of Education, National Center for Education Statistics' Education Longitudinal Study (ELS) of 2002 participants. In both the two- and four-year en-

rollment HGLM models, Promoting Power was found to significantly increase the likelihood of student matriculation, with the effect more pronounced for two-year enrollment.

176 Analysis of Statistical Models ■

Biopharmaceutical Section

Monday, August 1, 10:30 a.m.–12:20 p.m.

A Dose Finding Method In Joint Modeling For Mixed Type Of Efficacy And Safety Outcomes

✦Aiyang Tao, Novrtis Pharmaceutical, 75 westview rd, Short Hills, NJ 07078 USA, *aiyang_tao@yahoo.com*; Yong Lin, University of Medicine and Dentistry of New Jersey ; Jose Carlos Pinheiro, Johnson & Johnson PRD; Weichung Joe Shih, University of Medicine and Dentistry of New Jersey

Key Words: dose-finding, joint model, mixed outcomes

Determination of appropriate dose(s) to advance into Phase III is one of the most challenging and important decisions made during drug development. In clinical trials there are numerous cases that continuous and discrete endpoints are observed.It is very common that we need to consider joint continuous and discrete responses for the dose finding. This paper will address how to select dose(s) in Phase II trials by combining information about the efficacy and safety in a joint model setting for mixed type of outcomes. The methods we present in the paper may play a key role in drug development program and are often the gate-keeper for large confirmatory Phase III trials with greater chance for success of approval.

Inferences For The Poisson-Inverse Gaussian Distribution With Application To Multiple Sclerosis Clinical Trials

✦Mallikarjuna Rettiganti, University of Arkansas for Medical Sciences, Biostatistics Program, Slot 512-43, 1 Children's Way, Little Rock, AR 72202, *mrrettiganti@uams.edu*; Haikady Nagaraja, The Ohio State University

Key Words: Poisson Inverse Gaussian, likelihood ratio, score, Wald, multiple sclerosis

Magnetic resonance imaging (MRI) based new brain lesion counts are widely used to monitor disease progression in relapsing remitting multiple sclerosis (RRMS) clinical trials. These data generally tend to be over dispersed with respect to a Poisson distribution. It has been shown that the Poisson-Inverse Gaussian (P-IG) distribution fits better than the negative binomial to MRI data in RRMS patients selected for lesion activity during the baseline scan. In this paper we use the P-IG distribution to model MRI lesion count data from RRMS parallel group trials. We propose asymptotic and simulation based exact parametric tests for the treatment effect such as the likelihood ratio (LR), score and Wald tests. The exact tests maintain precise Type I error levels for small sample sizes when the asymptotic tests fail to do so. The LR test remained empirically unbiased and resulted in a 30-50% reduction in sample sizes required when compared to the Wilcoxon rank sum (WRS) test. One of the Wald tests had the highest power to detect a reduction in the number of lesion counts and provided a 40-57% reduction in sample sizes when compared to the WRS test.

Multiple Imputation To Correct For Covariate Measurement Error Based On Summary Statistics From External Calibration Data

✦Ying Guo, Merck & Co., Inc., , *ying.guo2@merck.com*; Rod Little, University of Michigan

Key Words: calibration data, measurement error, multiple imputation, regression calibration

Covariate measurement error is very common in empirical studies, and currently information about measurement error provided from calibration samples is insufficient to provide valid adjusted inferences. We consider the problem of estimating the regression of outcomes Y on covariates X and Z, where Y and Z are observed, X is unobserved, but a proxy variable W that measures X with error is observed. Data on the joint distribution of X and W (but not Y and Z) are recorded in a calibration experiment. The data from this experiment are not available to the analyst, but summary statistics for the joint distribution of X and W are provided. We describe a new multiple imputation (MI) method that provides multiple imputations of the missing values of X in the regression sample, so that the regression of Y on X and Z and associated standard errors are estimated correctly using multiple imputation (MI) combining rules, under normal assumptions. The proposed method is shown by simulation to provide better inferences than existing methods, namely the naÔve method and regression calibration, particularly for correction for bias and achieving nominal confidence levels.

Measurement Error On Cost-Effectiveness Ratio And Net Benefit

✦Ruifeng Xu, Merck Research Laboratories, Mail Stop 1C-60, 351 N Sumneytown Pike, North Wales, PA 19454, *ruifeng_xu@merck.com*; Ping-Shou Zhong, Iowa State University; John R. Cook, Merck Research Laboratories, North Wales, PA

Key Words: measurement error, cost-effectiveness, net benefit

Measurement error on cost-effectiveness analysis has received increasingly attention in the pharmaceutical industry. In this presentation, we explore the influence of measurement error on the cost-effectiveness ratio and net benefit. We found that the measurement error on the biomarker could induce sufficient large bias in the estimation of cost-effectiveness ratio. A deconvolution method and Simulation Extrapolation method are proposed to correct the bias. The asymptotic normalities of the new estimators are derived. In the second part, a realistic modeling method, multi-cycle patient level simulation model is explored to assess the impact of measurement error on the cost-effectiveness ratio and net benefit. Our results show that the measurement error could underestimate the net benefit and overestimate the cost-effectiveness ratio.

Shape-Restricted Regressions With Heteroscedastic Variances Subject To Order Constraints And Their Applications In Bioassay

✦Huitian Xue, University of Hongkong, Hongkong, None China, *mthgh123@live.com*

Key Words: shape restricted regression, rectangle constraints, likelihood function, restricted DP algorithm

Restricted parameter problems arise in many applications, for example, in ordinal regression, sample surveys, bioassay, dose-response, variance components models,and factor analysis models. To investigate the dose-response relationship of certain drug in developing, we often need to conduct a dose-response experiment with multiple groups associated with multiple dose levels of the drug. The dose-response relationship can be modeled by a shape-restricted normal regression. We develop EM-type algorithms to estimate normal means and variances subject to constraints simultaneously. These constraints include the simple order, simple tree order, umbrella order, and so on. Applications to the analyses of two real data on radioimmunological assay of cortisol and bioassay of peptides are presented to illustrate the proposed methods.

A Regression Model For General Trend Analysis Of Bivariate Continuous Panel Data

✦Wei-Hsiung Chao, National Dong Hwa University, Department of Applied Mathematics, No. 1, Sec. 2, Da Hsueh Road, Shoufeng, Hualien 97401, International Taiwan, *whchao@mail.ndhu.edu.tw*; Yi-Ran Lin, National Dong Hwa University

Key Words: Markov processes, longitudinal data, generalized estimating equations, Ornstein-Uhlenbeck processes

In many longitudinal studies, the underlying continuous response processes of interest may undergo natural fluctuation over time, as is the case with blood pressure. In these situations, transitions between specific states are often not as interesting as the general direction and rate of movement, since uncountably many transitions between states are involved and they do not provide a summary information about the evolution of the process. To provide such general trend information for panel data of a single continuous response, Chao and Chen (2009) developed a Markov-based regression model that can be viewed as a continuous time first order autoregressive regression model with time varying lag effects of the covariate. In this talk, we further extend their model to settings of panel data of two continuous response variables. In addition to assessing the general trend of each underlying response process, the proposed model can determine if there is feedback effect of one response process on the other process. For robust inference on parameters in the conditional mean structure, the generalized estimating equation approach is adopted.

177 Issues in Clinical Trial Design ■

Biopharmaceutical Section

Monday, August 1, 10:30 a.m.–12:20 p.m.

A Practical Guide To Drug Combination Models

✦WEI ZHAO, MEDIMMUNE, 1 MEDIMMUNE WAY, GAITHERSBURG, MD 20878, *ZHAOW@MEDIMMUNE.COM*; LANJU ZHANG, MEDIMMUNE; LINGMIN ZENG, MEDIMMUNE; HARRY YANG, MEDIMMUNE

Combination of drugs becomes more promising, especially in treating malignant cancers. Scientists are interested in identifying compounds that act synergistically when combined. There are many statistical methods proposed to estimate the combination index and to construct its confidence interval in the literature. Also made available are statistical software packages for evaluations of combination therapies. With extensive practical experiences in application of different models and statistical packages, we have modified some of the packages to make them both more user-friendly and robust under unconventional conditions. In this presentation, we will compare and demonstrate the strength and weakness of different methods and software packages.

The Price We Pay To Conduct Ethical Trials For Efficacy

✦Charlie Goldsmith, Simon Fraser University, Blusson Hall 9510, 8888University Drive, Burnaby, BC V5A 1S6 Canada, *charles_goldsmith@sfu.ca*

Key Words: factorial design, partial factorial design, efficacy, efficiency, ethics, patients

Previous work has shown a partial factorial design that eliminates those treatment combinations of unknown efficacy, such as in a trial of a new therapy where the clinical condition has at least two already proven therapies that are used alone or in combination. Such partial designs use fewer treatment combinations but all provide patients with therapy of known efficacy. Partial designs allow efficacy estimation of the unproven therapy, provided this new therapy does not interact with the other therapies. Relative to a full factorial design of 3 therapies at 2 levels in a 3 replicate 2**3 = 8 treatment combination design with 24 patients randomized, the partial factorial design with 6 treatment combinations eliminating the placebo group and the new therapy group, using 4 replicates with 24 patients, estimates new therapy efficacy with 80% efficiency because of 3 new therapy treatment effect replicates, rather than the 4 provided by the full factorial design. Partial designs allow for new therapy efficacy estimation when the other therapies are additive, synergistic or antagonistic, as long as the new therapy does not interact with the proven therapies.

Select Matching Controls For Treated Cohort

✦Xin Chen, Department of Pathology, UC Irvine, , *xinc6@uci.edu*; Zhenyu Jia, Department of Pathology, UC Irvine; Dan Mercola, Department of Pathology, UC Irvine

Key Words: classification, clinical

Control cohort was missing in a study of an adjuvant multimodality therapy for prostate cancer. We hope to select matching untreated patients from another study to complete the survival comparison and justify the efficacy of the therapy. This is a modified statistical classification problem where traditional classification methods such as logistic regression and support vector machine do not apply since only one group is specified in the training set. We proposed an unsupervised clustering method which involves the calculation of the weighted Euclidean distance based on clinical variables of interest, such as age, pre-OP PSA, Gleason score and etc. The control patients that are close to the treated cohort in terms of Euclidean distance will be selected. The method was verified by analyzing two simulated datasets, and further demonstrated using prostate cancer patient data.

Assessing The Causal Effect Of Treatment Dosages In The Presence Of Self-Selection

✦Xin (Cindy) Gao, University of Michigan, 1415 Washington Heights, Department of Biostatistics, Ann Arbor, MI 48109, *xingao@umich.edu*; Michael R. Elliott, University of Michigan

Key Words: causal modeling, potential outcome, principal stratification, missing data

To make drug therapy as effective as possible, patients are often put on an escalating dosing schedule. But patients may choose to take a lower dose because of side effects. Thus, even in a randomized trial, the dose level received is a post-randomization variable, and comparison with the control group may no longer have a causal interpretation. Hence we use the potential outcomes framework to define pre-randomization "principal strata" from the joint distribution of doses selected under control and treatment arms, with the goal of estimating the effect of treatment within the subgroups of the population who will select a given set of dose levels. When subjects on the control arm cannot obtain treatment, these principal strata are fully observed on treatment, but remain latent on control. Adverse event information can be used to identify the tolerated dose level in the control arm.

Information Divergence And The Evaluation Of Surrogate Markers In Clinical Trials

✦Xiaopeng Miao, Department of Biostatistics, Boston University, , *miaox@bu.edu*; Ashis Gangopadhyay , Boston University

Key Words: surrogate marker, information divergence

Recently, there is a growing interest in identifying surrogate markers in clinical trials for the purpose of reducing the cost of drug development and better understanding of the disease. Although surrogate markers are usually proposed based on biological considerations, the evaluations of surrogate markers depend largely on statistical methods. In this paper, we review existing methods for evaluating surrogate markers in single-trial framework and discuss the strengths and limitations of these methods under various settings. We also propose new methods to assess surrogate markers that are based on various information divergence criteria between nested models, and provide comparisons of the performances of these proposed new measures with the existing methods.

Modified Zelen'S Approach Randomization In Studies With Unequal Allocation

✦Olga M Kuznetsova, Merck Sharp & Dohme Corp., 126 E. LINCOLN AVENUE, Rahway, NJ 07065-0900, *olga_kuznetsova@merck.com*; Yevgen Tymofyeyev, Merck Sharp & Dohme Corp.

Key Words: Randomization, modified Zelen's approach, unequal allocation, multi-center study, dynamic allocation, stratified allocation

Modified Zelen's approach [Zelen 1974; McEntegart 2008] is a randomization technique useful in multi-center trials where balance in treatment assignments within a center is desired. It has great balancing properties in a study with equal allocation to several treatment arms [Morrissey et al., 2010]. This technique can also be used in a study with unequal allocation, where it would provide an allocation ratio close to the targeted one within centers as well as across centers - or across strata, if the allocation is stratified by factors other than center. However, the implementation of the modified Zelen's approach for unequal allocation involves more than just imposing a constraint on within-center imbalance in treatment assignments, as is the case with equal allocation. The presentation will explain why, and offer an easy way to expand the modified Zelen's approach to unequal allocation. The balancing properties of the modified Zelen's approach for unequal allocation, including stratified allocation and hierarchical dynamic allocation schemes that include modified Zelen's approach at the center level, will be examined through simulations.

178 Statistical Methods for High Dimensional Data ■●

ENAR, Section on Statistical Computing

Monday, August 1, 10:30 a.m.–12:20 p.m.

A Regularization/Extrapolation Corrected Score Method for Nonlinear Regression Models with Covariate Error

✦David Zucker, Hebrew University, , *mszucker@mscc.huji.ac.il*; Malka Gorfine, Technion - Israel Institute of Technology; Yi Li, Dana-Farber Cancer Institute and Harvard School of Public Health ; Donna Spiegelman, Harvard School of Public Health

Key Words: Errors in variables, nonlinear models, logistic regression

Many regression analyses involve explanatory variables that are measured with error, and ignoring this error leads to biased estimates for the regression coefficients. We present a new general method for adjusting for covariate error. Our method consists of an approximate version of the Stefanski-Nakamura corrected score approach, using the method of regularization for approximate solution of integral equations, along with an extrapolation device similar in spirit to that of the SIMEX method. Specifically, we compute estimates for various values of the regularization penalty parameter and extrapolate to a penalty parameter of zero. We develop the theory in the setting of classical likelihood models, covering nonlinear regression, logistic regression, and Poisson regression. The method is extremely general in terms of the types of measurement error models covered, and is a functional method in the sense of not requiring information on the distribution of the true covariate. We present a simulation study in the logistic regression setting, and provide an illustration on data from the Harvard Nurses' Health Study concerning the relationship between physical activity and breast cancer.

A New Model-Free Sure Independence Screening For Ultra-High Dimensional Problems

Runze Li, Penn State University; ✦Wei Zhong, The Pennsylvania State University, 333 Thomas Building, Department of Statistics, Penn State, University Park, PA 16802, *wxz123@psu.edu*; Liping Zhu, Shanghai University of Finance and Economics

Key Words: Variable Selection, Sure Screening Property, Ranking Consistency, Distance Correlation, Ultrahigh Dimensionality, Dimension Reduction

High dimensional regression analysis has become increasingly important in diverse fields of scientific research. In this paper we introduce a new model-free sure independent screening procedure to select important predictors when p >> n. It also allows us consider independent screening for group-wise predictors and multivariate responses. This proposed procedure imposes little assumption on regression structure, so it also allows arbitrary regression relationship between y and x. For theoretical properties, we demonstrate that the proposed independent screening procedure enjoys the ranking consistency property, that is, it can rank important predictors in the top consistently even when p>>n. Meanwhile, under some mild conditions, it has the sure screening property, that is, with a proper threshold, it can select all important predictors with probability approaching to one as n goes to infinity. In addition, a corresponding iterative procedure is proposed to enhance its finite sample performance. Numerical examples through comprehensive simulations and an application indicate that the new proposal performs quite well in a variety of ultrahigh dimensional regressions.

Integrating Data Transformation In Principal Components Analysis

✦Mehdi Maadooliat, Texas A&M University, College Station, TX 77840, *madoliat@stat.tamu.edu*

Key Words: Profile likelihood, Transformation model, PCA, Fuctional PCA, Missing data

Principal Component Analysis (PCA) is one of the commonly used dimension reduction techniques. However, PCA does not work well when there are outliers or the data distribution is skewed. One popular solution is to transform the data to resolve this abnormal behavior caused from skewness or presence of outliers. Usually, such transformations can be obtained based on extensive data exploration, previous studies, or prior knowledge of expertise. In this work, we present an automatic procedure to achieve this goal based on a statistical model with extensions for handling the missing data and functional data structure. The proposed technique transforms the data to vanish the skewness of the data distribution and simultaneously perform the standard PCA to reduce the dimensionality. Our method is cast into a profile likelihood framework for efficient computation.

Bias Of The Out-Of-Bag (Oob) Error For Random Forests

✦Matthew Mitchell, Metabolon, Inc., NC , *mwmitch2@aol.com*

Key Words: random forest, out of bag error, metabolomics

Random Forest is an excellent classification tool, especially in the -omics sciences such as metabolomics where the number of variables is much greater than the number of subjects, i.e., "n < < p." However, the choices for the arguments for the random forest implementation are very important. With the default arguments, the out-of-bag (OOB) error overestimates the true error, i.e., the random forest actually performs better than indicated by the OOB error. This bias is greatly reduced by sampling without replacement and choosing the same number of samples from each group. However, even after these adjustments, there is a low amount of bias. The remaining bias occurs because when there are trees with equal predictive ability the one that performs better on the in-bag samples will perform worse on the out-of-bag samples. Cross-validation can be performed to reduce the remaining bias.

A Simulation Study To Evaluate The Operating Characteristics Of A Two-Stage Study To Develop And Validate A Panel Of Biomarkers For Predicting Prostate Cancer Recurrence

✦Joseph Koopmeiners, Division of Biostatistics, University of Minnesota, A460 Mayo Building, MMC 303, 420 Delaware St. SE, Minneapolis, MN 55455, *koopm007@umn.edu*; Rachel Rachel Isaksson Vogel, Biostatistics Core, University of Minnesota Masonic Cancer Center

Key Words: biomarkers, predictive model, two-stage design, ROC curve

We consider study design for a two-stage study to develop and validate a panel of biomarkers for predicting prostate cancer recurrence. In stage one, a predictive model for prostate cancer recurrence is developed using a set of candidate biomarkers. The study will be allowed to terminate for futility after stage one if initial estimates of prognostic accuracy are unacceptable. If initial estimates of prognostic accuracy are promising, the prognostic accuracy of the predictive model is evaluated using a set of independent samples in stage two. We present results from a simulation study to evaluate the effect of design parameters (the proportion of samples used in stage one and the cutoff for early termination) and marker parameters (number of markers truly associated with prostate cancer recurrence, correlation between markers, etc.) on the type-I error rate, power and expected sample size.

Financial Networks With The Graphical-2-Lasso

✦Alan Burton Lenarcic, UNC Chapel Hill, 5 Howell Street #7, Chapel Hill, NC 27514 USA, *alanjazztenor@gmail.com*

Key Words: Bayes, networks, lasso, finance, correlation, algorithms

Advances in network and model selection motivated by L1 convex algorithms now provide penalized likelihood estimators to study the realm of partial correlation and conditional independence in complex networks. The 2Lasso method was motivated to improve Lasso false positive rate with negligible additional computational complexity, giving Lasso an informative Bayesian latent data model that solves the question of noise coefficient thresholding. Thus 2Lasso is also applicable to graphical-lasso network inspection, and provides structure to avoid issues like correlation bias, that could hurt glasso in an arena like portfolio allocation, where accurate correlation estimates are essential. Incidences of financial contagion, or spikes in asset correlation, suggest moments where investors might need improved regularized covariance estimation. Anomalies like the May 5th 2010 crash, driven in part through algorithmic trading, show that these decisions need to be made quickly, often in milliseconds. Lasso's robustness, speed, well documented k < n < p performance, make it a promising candidate; oddly enough it can serve up Bayes confidence measures too!

Differential Bias Introduced By The Copy Number Variation Inheritance Model

✦Sulgi Kim, University of Washington, 4333 Brooklyn Ave. NE, Seattle, WA 98105, *sulgik@uw.edu*; Ellen M. Wijsman, University of Washington; Debby W Tsuang, University of Washington

Key Words: CNV, Genetics, Family data

Copy Number Variation (CNV) calling for SNP chip data is challenging. Wang et al. (2008) proposed a Hidden Markov Model for family data that jointly models all members of a trio (parents and their offspring) using Mendelian Inheritance Laws, with distribution of an implementation in the package, "PennCNV Joint-calling". Conceptually, this formal model is expected to improve CNV-detection in related samples because it uses more information, which the authors showed with simulated datasets. We show that CNVs inferred by this inheritance model have unequal marginal sensitivity between a parent and an offspring thus introducing a differential bias. Therefore, a typical family study where the phenotype is correlated with the family membership (e.g. the case-parent design) violates the assumption that case and control should have errors equally. This results in an inflated CNV detection rate was inflated in the offspring in our real data. We show the bias of the model theoretically, with simulation, and then with two real datasets.

179 Nonlinear and Nonstationary Time Series

Business and Economic Statistics Section

Monday, August 1, 10:30 a.m.–12:20 p.m.

Bispectral-Based Methods For Clustering Nonlinear Time Series

✦Bonnie K Ray, Business Analytics and Math Sciences, IBM T.J. Watson Research, PO Box 218, Yorktown Heights, NY 10598, *bonnier@us.ibm.com*; Jane Harvill, Baylor University; Nalini Ravishanker, University of Connecticut

Key Words: nonlinear, clustering, frequency domain, bispectrum

It is well-known that in general, second-order properties are insufficient for characterizing nonlinear time series. In particular, the normalized bispectral density function is constant for a linear, Gaussian series, but typically not for nonlinear series. Furthermore, different nonlinear time series models have different bispectral signatures. Based on these properties, we propose the use of distance measures based on the squared modulus of the estimated normalized bispectrum as a means for clustering nonlinear series. In this talk, we summarize the performance of agglomerative clustering methods that use distance measures computed from the estimated bispectrum for a mix of linear and nonlinear time series. We then present an application of the methods to a set of gamma-ray burst time profiles, to aid astrophysicists in identifying sets of gamma-ray bursts emanating from the same type of astral event.

Segmenting Nonstationary Time Series Via Quantile Autoregressions

✦Ming Zhong, University of California, Davis, 2900 Solano Park Circle #3424, Davis, CA 95616, *mgzhong@ucdavis.edu*

Key Words: quantile autoregression, break point, minimum description length, genetic algorithm

Many time series display non-stationarities, especially if data is collected over long time spans. Since parameter estimates and forecasts can be severely biased if non-stationarities are not taken into account, identifying and locating structural breaks has become an important issue. In practice, one commonly observes error distributions with longer tails than that of the Gaussian distribution, and thus parameter estimates obtained from an application of the Gaussian likelihood may be inefficient. To incorporate skewed and possibly heavy-tailed innovations into the model fitting, we propose the use of quantile autoregression models with Asymmetric Laplace innovations. In this setting, we try to detect structural breaks with the minimum description length principle selecting the number and locations of break points for nonstationary time series through genetic algorithm. Large sample properties and theoretical justifications for the consistency are provided, and numerical results from simulations and data applications show that our method consistently estimates the number and locations of the breaks. (Joint work with Alexander Aue and Thomas Lee.)

Revisiting Levene'S Statistic For Change Point Detection In Variance And Its Application To Dow-Jones Average Index

✦Kyungduk Ko, Boise State University, 1910 University Dr., Department of Mathematics, Boise, ID 83725-1555, *ko@math.boisestate.edu*

Key Words: Dow-Jones Average, Levene statistic, Power of test, Null hypothesis distribution, Variance change point

We propose a method to estimate an unknown variance change point in a sequence of observations. Levene's statistic, which is used for testing equality of population variances, is adopted for measuring a potential discrepancy between the variances of two subsequences before and after each observation. The maximum of those Levene's statistic values is then used as a test statistic for testing the null hypothesis of no change point. The null hypothesis percentiles of the proposed test statistic are obtained via Monte Carlo simulation, and using the critical points, empirical test sizes and powers are presented as well. Applications of the proposed method to the weekly closing values of the Dow-Jones Industrial averages are given.

Breaking Trends And The Prebisch-Singer Hypothesis: A Further Investigation

✦MOHITOSH KEJRIWAL, PURDUE UNIVERSITY, 403 WEST STATE STREET, ROOM 410, KRANNERT SCHOOL OF MANAGEMENT, PURDUE UNIVERSITY, WEST LAFAYETTE, IN 47906 USA, *mohitoshk@gmail.com*

Key Words: structural breaks, trend functions, Prebisch-Singer Hypothesis, unit roots, primary commodity prices

This paper examines the Prebisch-Singer Hypothesis employing new time series procedures that are robust to the nature of persistence in the commodity price shocks, thereby obviating the need for unit root pretesting. Specifically, the procedures allow consistent estimation of the number of structural breaks in the trend function as well as facilitate the distinction between trend breaks and pure level shifts. In comparison with past studies, we find fewer cases of commodities that display negative trends thereby weakening the case for the Prebisch-Singer Hypothesis. Finally, a new set of powerful unit root tests allowing for structural breaks under both the null and alternative hypotheses is applied to determine whether the underlying commodity price series can be characterized as difference or trend stationary processes. Relative to the extant literature, we find more evidence in favor of trend

stationarity suggesting that real commodity price shocks are mostly of a transitory nature. This paper is based on joint work with Atanu Ghoshray (University of Bath) and Mark Wohar (University of Nebraska-Omaha).

Interrupted Cointegration With An Application To International Contagion

✦Luis Filipe Martins, ISCTE-LUI, Business School, Lisbon, Portugal, *luis.martins@iscte.pt*; Vasco Gabriel, University of Surrey

Key Words: Cointegration, Local Nonstationarity, Markov Switching, Financial Markets

We propose a new class of single-equation cointegration models in which the long-run equilibrium relationship is momentarily interrupted. This paper is closely related to Kim's specification (2003, Inference on segmented cointegration, Econometric Theory) . However, unlike `segmented' cointegration, our `interrupted' cointegration specification allows for (possibly) several short periods out of equilibrium. In addition, the statistical law behind the generating break points is stochastic. The equilibrium term follows a AR(1) model with an unobserved state process that is a stationary first-order Markov Chain in two states: stationarity and non-stationarity. Sufficient conditions for the existence of a strictly and second-order stationary solutions are provided. The estimation and inference technique is a two-stage procedure combining least squares and maximum likelihood methods. Lastly, we apply our proposed `interrupted' cointegration specification to the analysis of contagion in selected financial markets (USA, UK and Hong Kong). Using weekly data from March 1989 to April 2004, we find evidence of interrupted equilibrium around 1993 and between 1996 and 2000.

Asymptotic Behavior Of The Dickey-Fuller And The Augmented Dickey-Fuller Statistics Under Spurious Logarithms

✦Kalidas Jana, University of Texas at Brownsville, Department of Business Administration, 80 Fort Brown, Brownsville, TX 78520, *kalidas.jana@utb.edu*

Key Words: Dickey-Fuller statistic, Augmented Dickey-Fuller statistic, spurious logarithms

Inspired by "Spurious logarithms and the KPSS statistic," [Robert M. de Jong and Peter Schmidt, Economics Letters, 2002, 383-391], in this paper we investigate how the limiting distributions of the Dickey-Fuller and the Augmented Dickey-Fuller statistics are affected if the true time series is I(1) in level but logarithm has been spuriously applied to it.

Evaluating Garch Models Via Kalman Filter

✦Natalia Bahamonde, Pontificia Universidad Catolica de Valparaiso, Blanco Viel 596, Cerro Baron, Valparaiso, International CHILE, *natalia.bahamonde@ucv.cl*; Sebastian Ossandon, Pontificia Universidad Catolica de Valparaiso

Key Words: Time Series, GARCH process, extended Kalman Filter

In this work, we propose a novel estimation procedure for nonlinear time series models based on the extended Kalman filter. A popular nonlinear time series model, used in finance studies, is the GARCH models. The estimation of parameters in GARCH models is commonly performed using maximum likelihood procedure however when there is a limited amount of data or in the presence of missing data, this method may not be easy to apply. We show in this work that for GARCH processes, using the untransformed observations, it is possible to have a state space formulation and an efficient approach based on the extended Kalman filter in order to obtain the estimates of the parameters and the predictions. The extended Kalman filter proposed for GARCH models is derived from a discrete nonlinear state space formulation of the studied nonlinear model. Our method will be illustrated with a simulation study where we will compare our proposed method with standard estimation procedures specially in missing data situations because the Kalman filter techniques can handle missing observations. An application to real data set involving stock index is also reported.

180 Theory in Stochastic Processes and Graphs

IMS

Monday, August 1, 10:30 a.m.–12:20 p.m.

Power Of Roy'S Largest Root Test In Testing For Signal In Noise

✦Iain M Johnstone, Stanford University, Department of Statistics, 390 Serra Mall, Stanford University, CA 94305 USA, *imj@stanford.edu*; Boaz Nadler, Weizmann Institute of Science

Key Words: signal detection, Roy's largest root test, Matrix perturbation, Inverse Wishart distribution, MANOVA

We consider the problem of detecting the presence of a signal embedded in noise. It is assumed that we are given both signal bearing and noise-only samples, with Gaussian data in both samples. The problem reduces to either testing the equality of covariance matrices or testing for null regression against an alternative of a rank one perturbation. Using a matrix perturbation approach, combined with known results on the eigenvalues of inverse Wishart matrices, we study the behavior of the largest eigenvalue of the relevant matrix, and derive an approximate expression for the power of Roy's largest root test. The accuracy of our expressions is confirmed by simulations.

Zero Crossings In Gaussian Long Memory Processes

✦Mathieu Sinn, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1 Canada, *msinn@cs.uwaterloo.ca*

Key Words: Zero Crossings, Gaussian Processes, Long Memory, Binary Time Series, Hurst Parameter, Estimation

Zero Crossings (ZCs) are widely used, e.g., in signal processing or for the modelling of binary time series. In this talk, I will discuss properties of ZCs in Gaussian processes with long memory. I will first present a simple method for the numerical evaluation of ZC covariances; then I derive asymptotic expressions for the variance of ZC counts. Simulation studies for Fractional Gaussian Noise show that the distribution of the ZC counts is highly non-normal when the Hurst parameter is close to 1. At the end of my talk, I will demonstrate how the ZC counts in aggregated processes can be used to estimate the asymptotic index of self-similarity.

The Limit Distribution Of The Maximum Increment Of A Random Walk With Dependent And Regularly Varying Jump Size

✦Martin Moser, Technische Universit‰t M¸nchen, Zentrum Mathematik, Lst. f. math. Stat. M4, Boltzmannstr. 3, Garching bei M¸nchen, International 85748 Germany, *moser@ma.tum.de*; Thomas Mikosch, University of Copenhagen

Key Words: change point analysis, extreme value theory, maximum increment of a random walk, regular variation, dependent random walk

For a sequence of regularly varying random variables X_i we consider the following statistical problem motivated by change point analysis: We test the null hypothesis of constant mean $E X_1 = \ldots = E X_n = \mu$ against the epidemic alternative of a change in the mean, i.e. $E X_1 = \ldots = E X_k = E X_{m+1} = \ldots = E X_n = \mu$ and $E X_{k+1} = \ldots = E_m = \nu$, where $\mu \not= \nu$ and $1 \leq k < m < n$. This leads to a test statistic T_n given by the normalized maximum increment of the random walk $S_n = X_1 + \cdots + X_n$, $S_0 = 0$.\par For statistical reasons, it is very importand to understand the limit distribution of T_n. In this talk, we will derive this limit distribution for different linear and nonlinear dependence structures of the random variables X_i. It will turn out that T_n will converge to a FrÈchet distribution in every case.

A Study Of Some Different Concepts Of Symmetry

✦Wen-Jang Huang, National University of Kaohsiung, No. 700, Kaohsiung University Rd. Nanzih District, Kaohsiung, International 811 Taiwan, R.O.C., *huangwj@nuk.edu.tw*; Hui-Yi Teng, National University of Kaohsiung

Key Words: characterization, doubly symmetry, I-symmetry, log-symmetry, R-symmetry, skewing representation

Recently, different concepts of symmetry on R+ such as R-symmetry, log-symmetry, and doubly symmetry are studied. Analogue concept and their properties of these symmetries on R will be studied in this work. Based on skewing representation and previous studies, characterizations of doubly symmetry on R will be given. Among others, some interesting examples of the so-called I-symmetry, which is the analogue of log-symmetry on R, will also be presented.

Estimation Of Fractal Dimension Of A Gaussian Field Via Euler Characteristic

✦Khalil Shafie, Shahid Beheshti University, Iran, *khalil.shafie@unco.edu*; Alireza Taheriyoun, Shahid Beheshti University

Key Words: Differentiable field, Euler characteristic, Fractal, Fractal dimension, Fractal index, Isotropic field

In working with real-valued Gaussian fields, one of the main problem is measuring and comparing the roughness of the field. The fractal index and fractal dimension are the most useful characteristics to measure the roughness of a surface. In recent years, different methods have been developed for estimating the dimension from the observed surface. The purpose of this paper is to use the Euler characteristic of the excursion set of a smoothed version of a fractional Gaussian random field to estimate its fractal dimension. We show that the estimator is almost consistent and apply the results to a simulation study and a real waveguide surface.

On Graph Cut And The Cheeger Constant

✦Bruno Pelletier, Universite Rennes 2, Rennes, 35043 France, *bruno.pelletier@univ-rennes2.fr*; Ery Arias-Castro, Department of Mathematics, UCSD; Pierre Pudlo, Universite Montpellier 2

Key Words: Isoperimetric inequality, Graph cut, Clustering, U-statistics, Empirical processes

In connection with graph partitioning clustering algorithms, we consider the estimation of the sets minimizing the Cheeger constant of a subset M of R^d. The Cheeger constant minimizes the ratio of a perimeter to a volume among all subsets of M. Given an n-sample drawn from the uniform measure on M, we introduce a regularized version of the conductance of the neighborhood graph defined on the sample. Then, we establish the convergence of the regularized conductance to the Cheeger constant of M. In addition, we prove the convergence of the sequences of optimal graph partitions to the Cheeger sets of M for the topology of L1(M). This result implies the consistency of a penalized bipartite graph cut algorithm.

Walsh'S Brownian Motion On A Graph

✦Kristin Jehring, Saint Mary's College, Department of Mathematics, Saint Mary's College, Notre Dame, IN 46556, *kjehring@saintmarys.edu*

Key Words: Brownian motion, Markov chain, harmonic function, graph, reversibility

We examine a variation of two-dimensional Brownian motion introduced in 1978 by Walsh. Walsh's Brownian motion can be described as a Brownian motion on the spokes of a (rimless) bicycle wheel. We will construct such a process by randomly assigning an angle to the excursions of a reflecting Brownian motion from 0. With this construction we see that Walsh's Brownian motion in the plane behaves like one-dimensional Brownian motion away from the origin, but at the origin behaves differently as the process is sent off in another random direction. We generalize the state space to consider a process on any connected, locally finite graph obtained by gluing a number of planar Walsh's Brownian motion processes together. In this generalized situation, we classify harmonic functions. We introduce a Markov chain associated to Walsh's Brownian motion on a graph and explore the relationship between the two processes, specifically the reversibility of the two processes. We derive formulas for the transition probabilities of the so-called embedded Markov chain and for the passage times of the Walsh's Brownian motion.

181 Classroom Examples and Technique ■●

Section on Statistical Education

Monday, August 1, 10:30 a.m.–12:20 p.m.

Teaching Statistics Using The News Media

✦T. Paulette Ceesay, Merck and Comapny, 351 N. Sumneytown Pike, North Wales, PA 19454, *paulette_ceesay@merck.com*

Key Words: teaching, media, statistical thinking

In this age of the internet, society is continuously bombarded with information 24 hours a day. Recently, a lot of emphasis has been placed on teaching students statistical concepts and how to apply them to the understanding of the real world instead of regurgitating and manipulating statistical methods. This is a particularly valuable approach for teaching statistics to students in non-quantitative majors such as criminal justice and journalism who may suffer from mathematics anxiety. The media is an ideal medium to accomplish this task. The use of statistical thinking concepts to help students critically analyze information in the media in areas such as health, science, and politics will be discussed as well as the challenges of teaching statistics deemphasizing quantitative methods.

Making Statistics Memorable: New Mnemonics and Motivations

✦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: mnemonic, memory, language, visual, learning, aesthetic

Mnemonics (memory aids) have the potential to decrease anxiety and increase recall. They may also serve as triggers for thinking routines (e.g., Pfannkuch, 2010) and increase cognitive resources available for more conceptual thinking called for by the Guidelines for Assessment and Instruction in Statistics Education (ASA, 2005). While examples of mnemonics are familiar in mathematics (e.g., FOIL, PEMDAS, SOHCAHTOA), there appears to be far less awareness of mnemonics available for statistics class and the first paper on the topic appears to be only one year old (N. Hunt's 2010 paper in Teaching Statistics). This JSM paper discusses some examples from a longer 2011 paper by Lesser in Model Assisted Statistics and Applications. Examples vary in function (fact vs. process) and in form (letter-based, image, jingle). The functional and aesthetic qualities of the best examples may help make statistics memorable in every sense.

Statistical Humor: An Oxymoron?

✦Harry James Norton, Carolinas Medical Center, Research Office Building #410, P.O. Box 32861, Charlotte, NC 28232-2861, *jnorton@carolinas.org*

Key Words: Teaching, Statistical Education, Humor, Jeopardy style questions

Warning: this presentation contains no statistical theorems, proofs, or examples of innovative statistical procedures. Saul Bellow wrote in Herzog, "An utterly steady, reliable woman, responsible to the point of grimness, Daisy was a statistician for the Gallup Poll." Does Bellow's stereotype of statisticians ring true? Is the field without jokes, cartoons, statistical heroes or tales about statistics saving the day? Keeping undergraduate biology students and medical residents interested in statistics when the majority of the students are taking the class as a requirement can be challenging. I will present examples of jokes, stories, cartoons, poems, and one-liners about statistics, statisticians and teaching that I use in my classes to lighten the mood. I will pose Jeopardy style questions to the audience challenging them to identify quotes on the subject of statistics from famous people both real and fictional. Participants will learn why there is no Noble Prize for Mathematics and Statistics, discover who was the most evil mathematician of all time, and hear about the short story where a man uses his extraordinary knowledge of statistics to charm a woman into marriage.

Grammar of Association

✦Milo Schield, StatLit Project, 9 Robb Farm Road, North Oaks, MN 55127, *milo@pro-ns.net*

Key Words: statistical literacy, grammar, association, syntax, semantics, causation

The distinction between association and causation is central to statistics. Yet the grammar used to describe this difference often creates confusion. This paper reviews different kinds of associations and compares them with the various grammatical devices used to indicate an association.

Using Pedometers To Collect Data: Just How Accurate Is That Pedometer?

✦Phyllis Jane Curtiss, Grand Valley State University, 1 Campus Drive, Dept. of Statistics MAK A-1-178, Allendale, MI 49401, *curtissp@gvsu.edu*

Key Words: pedometers, activities, descriptive statistics, confidence interval, hypothesis testing

According to www.thewalkingsite.com, guidelines say we should walk 10,000 steps per day, which is close to 5 miles. A pedometer will keep track of the number of steps you take, but just how accurate is it? If the pedometer says 200 steps, did you really take exactly 200 steps? Does the brand of pedometer make a difference? I have developed activities that use pedometers in statistics classes for data collection and illustration of core statistical concepts. These activities help students see the usefulness of statistics in their everyday lives. They also help make students think about the benefits of physical activity. The activities give ideas on collecting data that can then be analyzed in class activities or on homework assignments. Activities that illustrate descriptive statistics, a confidence interval for the mean and the paired t-test will be discussed.

It May Be A Great Day For Baseball, But Is It A Great Day For A Knuckleball?

✦Robert H. Carver, Stonehill College, 320 Washington Street, Easton, MA 02357, *rcarver@stonehill.edu*

Key Words: Multiple regression, Baseball, Inference, Decision analysis

The knuckleball is one of the rarest pitches in the repertoire of major league pitchers. It has the potential to confound batters with its unpredictable movement, which results from the absence of rotation on the ball and the interplay of the air turbulence and pressure differentials on the stitches and smooth surface of the baseball. The pitch is notoriously difficult to control, but when effective its slow speed and wide arc leaves batters extremely frustrated. Given the crucial role of air pressure and movement, one wonders if atmospheric and climatological variation influence a skilled pitcher's ability to control the knuckleball in a given

game. This paper examines game day data for veteran knuckleballer Tim Wakefield of the Boston Red Sox and finds that at least part of the answer may be blowin' in the wind. The analysis is accessible to undergraduate students, illustrating the managerial utility of multivariate models.

An Argument For Teaching Metrology In Introductory Statistics Classes

✦Emily Casleton, Iowa State University, Department of Statistics, Iowa State University, Ames, IA 50011, *casleton@iastate.edu*; Amy Borgen, Mathematica Policy Research, Inc.; Ulrike Genschel, Iowa State University; Alyson Wilson, Iowa State University

Key Words: metrology, education, variability

Undergraduate students in introductory statistics courses often struggle with the concepts of variability and how statistics will translate to their lives beyond the classroom. The aim of this research is the use of metrology, the science of measurement, to increase the understanding of these difficult concepts. Measurement quality and the inherent variability introduced through the measurement process are under emphasized topics in the statistics curriculum. To this end, materials and methods have been developed for use in introductory statistics courses. This material explains how to characterize sources of variability in a data set which is natural and accessible because sources of variability are observable, i.e. device or operator. Everyday examples of measurements, such as the amount of gasoline pumped into a car, are presented and the consequences of variability within those measurements are discussed. These materials were implemented into an introductory statistics course at Iowa State University. Student's subsequent understanding of variability and attitude toward the usefulness of statistics were analyzed in a comparative study.

182 Some Statistical Analysis Topics: Categorical Data ■

Section for Statistical Programmers and Analysts, Section on Statistical Graphics, International Chinese Statistical Association

Monday, August 1, 10:30 a.m.–12:20 p.m.

On Information Analysis Of Large Categorical Data

✦Philip E Cheng, International Chinese Statistical Association, Institute of Statistical Science, Academia Sinica, Taipei, 115 TAIWAN, *pcheng@stat.sinica.edu.tw*; Keng-Min Lin, Academia Sinica; Juin-Wei Liou, Academia Sinica; Michelle Liou, Academia Sinica

Key Words: Linear Information Models, Log-linear Models, Model Selection, Mutual Information

Mutual information identities have provided an insight into statistical inference of categorical data. Recent studies by the authors showed that the fundamental likelihood structure presents information identities as linear information models (LIMs) directly observed from data. In contrast to the hierarchical log-linear models (LLM), LIM organize data information and select subsets of highly associated variables for useful dimension reductions out of a large multi-way contingency table. Parsimonious LIMs can be formulated and tested for the target variables of interest. An application of LIM is examined against a large and sparse banking credit-card risk categorical data of twenty-six variables, and various data interpretations are tested.

Marginal Models By Data Information

✦Michelle Liou, Academia Sinica, 115 TAIWAN, *mliou@stat.sinica.edu.tw*; Juin-Wei Liou, Academia Sinica; Philip E Cheng, International Chinese Statistical Association

Key Words: Mantel-Haenszel test, Marginal models, Linear Information Models, Differential item functioning

From the geometry of invariant Pythagorean laws (Stat. Sinica, 2008), a two-step likelihood ratio test (JASA, 2010) is used as a substitute for the commonly used Mantel-Haenszel test (J. Nat. Cancer Res., 1959). As a generalization, linear information models can be used to organize the information structures in a large scale categorical data analysis. A scheme for selecting associated variables for a parsimonious model is formulated and applied to an empirical data analysis. For diagnosis of dementia, individual cognitive abilities screening instrument (CASI) scales have been commonly used for identifying patients from normal subjects. Let normal subjects form a reference group and patients be the focal group, differential items functioning (DIF) of CASI subscales were assessed in a multi-way contingency table. Interactions and partial associations were tested using the information model approach. It effectively examines DIF between the CASI subscales and background variables.

Constrained Maximum Likelihood Estimate Of Cell Probabilities Under Local Odds Ratio Constraints

✦Hsun-chih Kuo, National Chengchi University, seankuo@nccu.edu.tw, Taipei, 11605 Taiwan, *seankuo@nccu.edu.tw*

Key Words: Constrained MLE, Local odds ratio Ordering, Lagrangian.

When natural odds ratio ordering exists in a contingency table, it is desirable to incorporate the natural ordering into estimatation of the cell probability. In this study, we intent to estimate the constrained MLE of the cell probabilities of a contingency table under local odds ratio constraints. We will consider four approaches: (1) convex optimization type of Lagrangian approach, (2) penalized Lagrangian approach modified from the first approach, (3) iterative algorithm that solves one local odds ratio constraint at a time, and (4) an approach to identify the ACTIVE local odds ratio constraints first and then utilize the closed-form solution to solve constrained MLE for these ACTIVE cells. The constrained MLE of the cell probabilities obtained from aforementioned four methods will be compared and discussed.

An Outlier-Robust Hierarchical Bayesian Auxiliary Model For High-Dimensional Binary Classification

✦Lai Jiang, Department of Mathematics, University of Saskatchewan, McLean Hall, 106 Wiggins Road, Saskatoon, SK S7N 5E6, Saskatoon, SK S7N 5E6 Canada, *laj773@math.usask.ca*; Longhai Li, Department of Mathematics, University of Saskatchewan

Key Words: high-throughput data, classification, outliers, Bayesian, hierarchical, MCMC

Probit Model and Logistic Model are widely used for classification problems in applied statistics. Both of them can be viewed as inserting a latent variable with a linear regression link, while the noise term is assumed from Gaussian/logistic distribution for analytical conveniences. However, in real world the sparsity of high dimensional data always intensify the outliers problem, where the assumptions of these methods fail and lead to non-robust results that are vulnerable to type 2 errors. In this work a hierarchical Bayesian auxiliary model that incorporates heavy-tailed and symmetric t distribution both for noise and regression parameters is proposed. By adopting this structure a computational feasible solution with sparse feature selection will be obtained via MCMC, with some stochastical and adaptive sampling methods embedded. We compare our model with both current popular methods(like LASSO, DLDA) and traditional auxiliary models. The results show that heavy-tailed and symmetric t distribution is more robust to non-Gaussian outliers.

Optimal Allocation For The Second Elementary Symmetric Function With Different Coefficients

✦CHIEN-YU PENG, Academia Sinica, Taipei, 11529 Taiwan, *chienyu@stat.sinica.edu.tw*

Key Words: Accelerated Tests, Elementary symmetric function, Optimal allocation, Quadratic Programming

In this paper we consider the problem of determining the optimal size allocation and optimal number of experimental conditions for the second elementary symmetric function with different coefficients. We derive analytical solutions for different cases of practical applications and use the general formulation to elucidate the foundation between different parametric models found in recent studies. A geometrical interpretation for the structure of some theoretical results will then be given. Our results enable more complex problems to be more tractable than the numerical search algorithms currently being used.

Non-Inferiority And Superiority Tests For Multiple Immunogenicity Endpoints In Vaccine Clinical Studies

✦Lihan Yan, FDA, MD , *lihan.yan@fda.hhs.gov*

Key Words: non-inferiority, superiority, multiple endpoints, vaccine, type I error

Non-inferiority and superiority of a new vaccine to an existing one in terms of immunogenicity in a vaccine study often involves multiple endpoints corresponding to multiple strains or serotypes. The statistical hypothesis is often constructed as an intersection-union test among multiple endpoints, which allows the hypothesis test to be done for each endpoint at the desired overall Type I error rate with the overall rate being still conservatively preserved. Following the closed testing principle, the Type I error for the superiority test performed subsequently to the non-inferiority test is also controlled without further adjustment. It is not uncommon that the investigators weigh in on success of a trial by individual hypothesis test results and make a declaration even though the success criterion constrained by the hypothesis is not met. This presentation explores the impacts on the Type I error rate control from a variety of scenarios where the claims are made deviating from the decision rule indicated by the hypothesis testing. These results may serve as a potential reference for the investigators as to cautions as well as levels of comfort during decision making.

Robust Estimation In Case Of Asymmetry

✦Xiaoian Xu, Brock University, 500 Glenridge Ave, St Catharines, ON L2S 3A1 Canada, *xxu@brocku.ca*

Key Words: Huber's function, location parameters, regression coefficients, contaminated distribution, relative efficiency

In this paper, we discuss the construction of a modified version of Huber's function for many situations that involve asymmetrically distributed data and/or that the underlying distribution is contaminated asymmetrically. This new function employs a measurement for asymmetry so that the unequally weighted contribution in the loss function can be associated with the degree of asymmetry. We investigate the robust estimation using our modified Huber's function for location parameters and for regression coefficients with various contaminated underlying distributions. Relative efficiencies are presented in comparing with least squared estimates and original Huber's estimates. A set of suitable tuning constants for this proposed function is attained. The results demonstrate that our function is efficient.

183 Methods for Spatial and Spatio-temporal Data ●

Section on Statistics and the Environment

Monday, August 1, 10:30 a.m.–12:20 p.m.

On Some Properties Of Generalized Linear Mixed Models For Spatial Count Data

✦Victor De Oliveira , University of Texas at San Antonio, One UTSA Circle, Department of Management Science and Statistics, San Antonio, TX 78249, *victor.deoliveira@utsa.edu*

Key Words: Geostatistics, Second-order properties, Spatial prediction

Spatial count data occur often in many of the earth sciences, but unlike spatial continuous data, few models are available in the literature for their analysis. Currently, the most commonly used model seems to be the spatial generalized linear mixed model, which is difficult to fit. Perhaps due to the latter, most of the recent literature has concentrated on computational methods to fit this model, and some of its basic model properties are not well understood. In this talk I explore some properties and limitations of this class of models, and illustrate the findings with simulated and real datasets.

Spatial Prediction Variance Estimation Using Generalized Degrees Of Freedom

✦Roberto Rivera, University of Puerto Rico, Mayaguez, PO Box 250330, Aguadilla, PR 00604, *roberto.rivera30@upr.edu*

Key Words: Spatial prediction variance estimation, Generalized degrees of freedom, mean squared prediction error

In practice rarely (if ever) is the spatial covariance known in spatial prediction problems. Often, prediction is performed after estimated spatial covariance parameters are plugged into the prediction equation. The estimated spatial association parameters are also plugged into the prediction variance of the spatial predictor. However, simply plugging in spatial covariance parameter estimates into the prediction variance of the spatial predictor does not take into account the uncertainty in the true values of the spatial covariance parameters. Therefore the plug-in prediction variance estimate will underestimate the true prediction variance of the estimated spatial predictor, especially for small datasets. We propose a new way to estimate the prediction variance of the estimated spatial predictor based on Generalized Degrees of Freedom using parametric bootstrapping. Our new estimator is compared to three other prediction variance estimators proposed in literature. The new prediction variance estimator sometimes performs best among the four prediction variance estimators compared.

Conditional Maximum Likelihood Estimation Of Spatial Variation In Disease Risk From Locations Subject To Geocoding Errors

✦Dale Zimmerman, University of Iowa, 2141 Brown Deer Rd, Coralville, IA 52241, *dale-zimmerman@uiowa.edu*

Key Words: Case-control data, Geocode, Positional accuracy, Relative risk, Spatial epidemiology

The accurate assignment of geocodes to the residences of subjects in a study population is an important component of the data acquisition/assimilation stage of many spatial environmental epidemiological investigations. Unfortunately, it is not a simple matter to obtain accurate point-level geocodes. Ignoring the positional errors in geocoded data may lead to biased estimators of spatial variation in disease risk. In this talk, we modify Diggle and Rowlingson's (1994) conditional maximum likelihood estimation procedure for spatial variation in disease risk from locations ascertained without positional error, so as to permit valid inferences to be made from locations observed with error. The performance of the modified method relative to the original is investigated by simulation, allowing us to address the question of how large the positional errors must be for the modified method to be worth the additional level of complexity.

Covariate-Based Parameterization Of Time-Varying Spatial Covariances

✦Daniel W. Gladish, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, *dwg512@mail.missouri.edu*; Christopher K. Wikle, University of Missouri; Scott H Holan, University of Missouri

Key Words: Bayesian Hierarchical modeling, Covariance parameterization, Spatio-temporal process, Exponential spectral representation, Environmental modeling

Environmental spatio-temporal modeling is commonly approached from a dynamic first-order (mean) perspective due to the typical assumption that the true underlying process is best thought of as spatial process evolving through time. Thus, these processes are dynamic in time with a spatial error structure. This spatial covariance structure may be affected by time varying factors exogenous to the mean structure of the model. As such, our modeling approach is to incorporate these factors within the parameterization of a time-varying spatial covariance structure. A concern resides in parameterizing these exogenous covariate factors in a manner such that the covariance structure remains valid. The exponential spectral representation developed by Bloomfield (1973) provides a parsimonious framework for such parameterization, allowing for no direct restrictions on the covariate parameters. Under this time-varying error covariance framework, we develop a spatio-temporal model within the Bayesian hierarchical state-space paradigm. We illustrate this methodology using examples from the lower-trophic ocean ecosystem.

Semiparametric Bayesian Model For Areal Data With Space-Time Varying Coefficients

✦Bo Cai, University of South Carolina, 800 Sumter Street, Suite 205, Columbia, SC 29208, *bcai@sc.edu*; Andrew B Lawson, Medical University of South Carolina; Monir Hossain, The University of Texas; Jungsoon Choi, Medical University of South Carolina

Key Words: Bayesian regression, hierarchical structure model, Dirichlet process, Space-time models

In spatial analysis, the effects of covariates on the outcome are usually assumed to be invariant across areas. However, the spatial configuration of the areas may potentially depend on not only the structured random intercept but also spatially varying coefficients of covariates. In addition, the distribution of spatially varying coefficients are not always normally distributed. In this case, the normality assumption could lead to potential biases of estimations. In this article, we propose a semiparametric space-time model from a Bayesian perspective. The spatially varying coefficients of space-time covariates are modeled by using the spatial Dirichlet process prior which yields data-driven deviations from the normality assumption. The proposed semiparametric approach evinces the improvement of prediction compared to usual Bayesian spatial-temporal models with normality assumption on spatial-temporal random effects and to the model with the spatial random intercept modeled nonparametrically. A simulation example is presented to evaluate the performance of the proposed approach with the competing models. A real data example is used for an illustration.

A Layered Approach To Spatio-Temporal Parameter Inference

✦Adam Jaeger, University of Georgia, 101 Cedar St, Athens, GA 30602, *apjaeger@uga.edu*; Lynne Seymour, The University of Georgia

Key Words: spatial, temporal, spatio-temporal, climate, crop yield

In order to make inference on a parameter estimated from spatially and temporally correlated data, we propose to model the parameter itself as a spatial variable. Using both simulated and actual data the distribution of the parameter will be determined by use of spatial modeling techniques on the temporal projection of the parameter in order to accurately make inference on the parameter. This technique will be used to make inference about the relationship between climatological factors and Canadian food crop yields.

A Comparison Of Spatial Prediction Techniques Using Both Hard And Soft Data

✦Megan (Liedtke) Tesar, University of Nebraska-Lincoln, 340 Hardin Hall North, East Campus-UNL, Lincoln, NE 68583, *megan.liedtke@doane.edu*; David Marx, University of Nebraska-Lincoln; Dr. Steve Kachman, University of Nebraska-Lincoln

Key Words: spatial, kriging, soft data, prediction, natural variables

There is often a large amount of variability surrounding the measurements of environmental variables. It is therefore important to develop tools which account for uncertain measurements (soft data) in addition to measurements with little or no variability (hard data). Traditional methods, such as ordinary kriging, however, do not account for an attribute with more than one level of uncertainty. Thus, a new method, called weighted kriging is proposed. This method was implemented and tested against two alternative kriging procedures. The first alternative used only the hard data and the second used both the hard and soft data but treated both as hard. Simulated case studies showed that weighted kriging consistently results in more desirable model fitting statistics.

184 Statistical Methodology in Environmental Epidemiology

Section on Statistics in Epidemiology, Section on Statistics and the Environment, Scientific and Public Affairs Advisory Committee

Monday, August 1, 10:30 a.m.–12:20 p.m.

Hierarchical Models For Parameter Estimation Of Seir Models With Application To Raccoon Rabies Spread

✦Gavino Puggioni, Emory University, Atlanta, GA 30322 United States, *gpuggio@emory.edu*; Luca Gerardo Giorda, Emory University; Leslie Real, Emory University; Lance Waller, Emory University

Key Words: Stochastic Differential Equations, SEIR models, Raccoon rabies spread, Hierarchical Models, Bayesian Estimation, Space time models

The goal of this paper is to propose methods for parameter estimation in SEIR models. We propose a direct lieklihood evaluation of a reduced version of a system of 4 differential equations system. Estimation is performed with MCMC techniques. The methodology is then applied to the raccoon rabies spread in the state of New York from 1990 to 2007. Data are reported cases collected at different townships across the state and simulated total raccoon populations.

Linking Unknown Source Exposures To Health Effects

✦Eun Sug Park, Texas Transportation Institute, The Texa A&M University System, 3135 TAMU, College Station, TX 77843-3135, *e-park@tamu.edu*

Key Words: Multiple air pollutants, Multivariate receptor modeling, model uncertainty

There has been increasing interest in assessing health effects associated with sources of air pollutants. One of the major difficulties with achieving this goal is that in most cases the pollution sources are unknown and source-specific exposures cannot be measured directly; rather, they often need to be estimated by decomposing ambient measurements of multiple air pollutants. This process, called source identification and apportionment, is challenging because of the issues of the unknown number of sources and non-identifiability. An approach to account for both model uncertainty caused by the unknown number of sources and non-identifiability and uncertainty in estimated source-specific exposures into evaluation of source-specific health effects is presented. The method will be illustrated on real mortality data and speciated PM2.5 data.

Analytic And Study Design Considerations When A Continuous Regression Outcome Is Assessed In Pooled Samples

✦Robert H Lyles, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, *rlyles@sph.emory.edu*; Amita K. Manatunga, Emory University; Emily Mitchell, Emory University

Key Words: Grouped data, Pooled data, Regression, Study design

The process of pooling biological samples prior to laboratory assessment has recently received renewed attention because of its potential cost efficiency implications for epidemiological studies. In this talk, we demonstrate how ordinary and weighted least squares are directly applicable to the linear regression analysis of continuous outcomes assessed in equal and unequal-sized pools, respectively. We then consider the use of subject-specific covariate information to obtain potentially marked efficiency gains for regression coefficient estimation, relative to random pooling. This informed pooling strategy is contrasted against the use of targeted sampling of subjects for individual laboratory assessment. As time permits, we will also demonstrate the feasibility of maximum likelihood analysis in a mixed linear model setting when repeated or longitudinal outcome data are pooled within subjects. These results make interesting case studies for students of linear models, while also providing insights that carry over to other types of regression modeling when pooling is employed.

Identifying The Most Harmful Chemical Components Of Fine Particulate Air Pollution In The Us: A Bayesian Variable Selection Approach

✦Yiping Dou, Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115, *ydou@hsph.harvard.edu*; Christopher Barr, Harvard University; Yun Wang, Harvard School of Public Health; Roger D. Peng, Johns Hopkins School of Public Health; Francesca Dominici, Harvard School of Public Health

Key Words: Bayesian stochastic search variable selection, Cardiovascular disease, Particulate matter, Chemical components, Air pollution, Poisson regression

We present a Bayesian variable selection approach to estimate the toxicity of a complex mixture of fine particulate air pollution associated with cardiovascular disease. We use a national database during 1999-2008 of hospital emergency admissions for cardiovascular disease, ambient levels of major fine PM chemical constitutes and meteorology

data across 119 U.S. counties. We identify the most harmful chemical components associated with cardiovascular hospital admissions for more than 12 million elder Medical enrollees.

Do More Accurate Exposure Predictions Always Improve Health Effect Inference?

✦Adam Szpiro, University of Washington, , *aszpiro@u.washington.edu*; Christopher Joseph Paciorek, Department of Biostatistics, Harvard School of Public Health; Lianne Sheppard, University of Washington

Key Words: Environmental Epidemiology, Prediction Modeling, Spatial Statistics, Measurement Error, Air Pollution

A unique challenge in air pollution cohort studies and similar applications in environmental and occupational epidemiology is that exposure is not measured directly at subject locations. Instead, pollution data from monitoring stations at different locations than the subjects are used to predict exposures based on a regression model that may also incorporate spatial smoothing. These predicted exposures are used to estimate the health effect parameter of interest. It has been widely assumed that minimizing the error in predicting the true exposure is desirable to improve efficiency. We show in a simulation study that this is not always the case, and we interpret our results in light of recently developed theory for measurement error with spatially misaligned data.

Assessing The Health Impact Of Multiple Environmental Chemicals: Pcbs And Hypertension In The National Health And Nutrition Examination Survey

✦Krista Christensen, US EPA, 20460, *Christensen.Krista@epa.gov*; Paul White, US EPA

Key Words: Hypertension, PCBs, NHANES

Classical epidemiological methods may not be suitable for evaluation of complex exposures. We used data from the 1999-2004 National Health and Nutrition Examination Survey to illustrate an approach to complex exposures, using as our example the association between serum PCB levels and risk of hypertension. First, unconditional logistic regression was used to estimate odds ratios (ORs) and associated 95% confidence intervals (CIs), controlling for potential confounding covariates. Next, correlation and multicollinearity among PCB congeners was evaluated, and clustering analyses performed to determine groups of related congeners. Finally, an optimally weighted sum was constructed to represent the relative strength of association for each congener. PCB serum concentrations were on average higher among those with hypertension. However, multivariate analyses showed mixed findings; total PCBs were not associated with risk, but clustering analyses identified groupings of similar PCBs and most informative PCBs. Using a weighted sum approach to equalize different ranges and potencies, PCBs 66, 118 and 187 were most strongly associated with increased risk of hypertension.

Integrated Latent Variable Modeling Of Air Pollution Effects On Respiratory Health In The Children'S Health Study

✦Sandrah P Eckel, University of Southern California, Los Angeles, CA 90089, *eckel@usc.edu*; Kiros Berhane, University of Southern California; Duncan Thomas, University of Southern California

Key Words: Latent variable modeling, Bayesian modeling, Environmental epidemiology

Air pollution has important public health impacts, particularly in susceptible subgroups such as children. Data from the Southern California Children's Health Study (CHS) motivates an integrated model, where the goal is to quantify the association of short- and long-term exposure to air pollution with a respiratory health outcome, while taking into account information on a biomarker (B) of airway inflammation, namely exhaled nitric oxide. We use a latent variable approach to tie together the outcome and the biomarker and employ Bayesian and Frequentist methods to compare the "net effect" of the exposure (E) on the outcome (Y) in the integrated model to the analogous value in a standard regression of Y on E without using B. We show that the latent variable model is identifiable and find that the two estimators are equivalent, but the integrated model provides additional insights over the standard approach by quantifying hypothesized pathways. Finally, we apply an integrated model to data in the CHS.

185 Invited Oral Poster Presentations

Biometrics Section, ENAR, Section on Bayesian Statistical Science, Section on Statistical Computing, Section on Statistics in Epidemiology, Section on Survey Research Methods

Monday, August 1, 10:30 a.m.–12:20 p.m.

Haplotype-Based Association Studies Under Complex Sampling

✦Daoying Lin, University of Texas at Arlington, 411 South Nedderman Drive, 478 Pickard Hall, Arlington, TX 76019, *daoying.lin@mavs.uta.edu*; yan li, University of Texas at Arlington

Key Words: Complex sampling, Weighted EM algorithm, Haplotype, Genetic association studies, Weighted logistic regression, Case control studies

A widely used design strategy in the study of haplotype-based genetic association is case-control studies. Numerous methods have been proposed to infer haplotypes and investigate the role of genetic variants in common diseases under simple random sampling (SRS). It is becoming common that complex sampling, which usually involves complexities like differential weighting and clustering, is utilized in genetic association studies. In this article, we have formalized a two-step prospective approach that applies to case control studies with complex sampling. At first step, we develop a weighted EM (WEM) algorithm to infer haplotype frequencies from unphased genotype data. At second step, we study the association of haplotypes and diseases by weighted logistic regression. Monte Carlo simulation studies are used to evaluate the performance of proposed method. It has been found that methods developed under SRS give biased estimates and overstate the significance level. In contrast, our method performs well, produce consistent estimates and maintain its significance level.

A Bayesian Model for Inference on Population Proportions

✦Raymond Okafor, University of Lagos, Department of Mathematics, Akoka-Yaba, Lagos, Nigeria, *okaforray@yahoo.com*; Ugochukwu Ahamefula Mbata, University of Lagos

Key Words: Estimation, Empirical Bayes, Faculties at Main Campus, Proportions using Vehicles, Sample Surveys, Traffic Congestion, Undergraduates, UNILAG, Use of Motor Vehicles.

Traffic congestion and parking difficulties, have become a major concern to members of the University of Lagos (UNILAG). UNILAG has witnessed unprecedented growth in student's enrolment, in the last ten years or so culminating in the current total enrolment of more than thirty-five thousand students, of which about twenty-five thousand are undergraduates. In order to study the worrisome traffic situation, independent, though similar, sample surveys of undergraduates of the eight faculties on the main campus were conducted in 2007. The purpose of the surveys was to collect data on undergraduates who owned or used motor vehicles on campus. Further, to investigate possible temporal trends, the surveys were repeated in 2009. The types of data obtained from the surveys provided avenue for the application of Empirical Bayes (EB) analysis to estimate the proportions of students of individual faculties who used motor vehicles. The main result is that in 2007 about one in four students used motor vehicles, and this result held almost across the faculties. Although results of the 2009 surveys were similar there were faculties that recorded some reduction in the estimated proportions.

Significance Analysis And Statistical Dissection Of Variably Methylated Regions

✦Andrew Jaffe, JHSPH, 615 N Wolfe St, Room E3011, Baltimore, MD 21205 US, *ajaffe@jhsph.edu*

Key Words: variably methylation regions (VMRs), bump finding, multiple testing, functional data analysis, preprocessing

It has recently been proposed that variation in DNA methylation at specific genomic locations may play an important role in the development of complex diseases such as cancer. Here we develop one- and two-group multiple testing procedures for identifying and quantifying regions of DNA methylation variability. Our method is the first genome-wide statistical significance calculation for increased or differential variability, as opposed to the traditional approach of testing for mean changes. We apply these procedures to genome-wide methylation data obtained from biological and technical replicates and provide the first statistical proof that variably methylated regions exist and are due to inter-individual variation. We also show that differentially variable regions in colon tumor and normal tissue show enrichment of gene regulating gene expression, cell morphogenesis, and development, supporting a biological role for DNA methylation variability in cancer.

Multi-Domain Composite Transform Modeling Of Nonstationary Time Series Data With High Frequency Content Using Bayesian Functional Mixed Models

✦Josue Guillermo Martinez, Texas A&M Health Science Center, 1266 TAMU, College Station, TX 77843, *jgmartinez@srph.tamhsc.edu*

Key Words: Bayesian Modeling, Data Registration, Functional Mixed Models, Data Transformation, Spectrogram, Non-stationary Time Series

We propose a general strategy for modeling and performing inference on non-stationary time series with high frequency content. The strategy involves mapping the original data via a two-stage transformation to a space where modeling is more amenable. The two-stage transformation involves a mapping from the time domain to the spectrogram domain, where results are more interpretable, followed by a second mapping to the wavelet domain, where flexible, parsimonious modeling is executed using Bayesian functional mixed models. This is an application of a general approach we call the multi-domain, composite transform functional mixed modeling (MDCT-FMM). We describe this modeling strategy and use it to characterize bat chirps which are represented by time series of auditory calls. The bats of interest are Brazilian free-tailed bats (Tadarida brasiliensis), which use a variety of vocalizations to communicate in many contexts, one of which is mating. We study the chirp syllable, derived from mating type songs, and look for systematic differences between groups of bats captured in two different locations in the central Texas region.

Genome-Wide Epistasis Screening For Asthma Associated Traits

✦Elena S Gusareva, Systems and Model, Montefiore Institute, University of LiËge; GIGA-R, University of LiËge, Sart-Tilman, Bldg. B28, LiËge, B-4000 BEL, *egusareva@ulg.ac.be*; Jeroen S Huyghe, Department of Biomedical Sciences, University of Antwerp; Elena S Gusareva, Systems and Model, Montefiore Institute, University of LiËge; GIGA-R, University of LiËge

Key Words: eosinophil counts, asthma genes, genome-wide association study, gene-gene interactions

Genome-wide association (GWA) studies of asthma and associated traits have identified numerous genes. A substantial portion of the heritability of these traits remains unexplained. Some variants, not detectable via main effects GWA study may manifest themselves only in interaction with other variants. To search for interacting genes involved in regulation of asthma associated traits (total IgE, eosinophils, FEV1, FVC, FEV1/FVC) we performed GWA epistasis screening in two family groups of asthma patients:CAMP (Childhood Asthma Management Program:814 cases and 467 trios) and CARE (Childhood Asthma Research and Education:796 cases and 338 trios) [dbGaP accession number phs000166.v1.p1.c1]. Individuals were genotyped with the Aymetrix 6.0 array. After quality control 574922 and 575010 SNPs in CAMP and CARE respectively, were tested with FBAT. No main effects genome-wide significant associations were found. We prioritized candidate pairs of SNPs for MB-MDR epistasis screening using Biofilter leading to 7632 SNPs for CAMP and 7603 SNPs for CARE. The most significant pair-wise interaction was identified between SNPs from loci 7p21.1 and 12q23.3 influencing eosinophil level in asthmatics.

A New Method For Detecting Associations With Rare Variants For Complex Disease

Huann-Sheng Chen, National Cancer Insitute; ✦Shunpu Zhang, University of Nebraska-Lincoln, , *szhang3@unl.edu*

Key Words: genetics, complex disease, association, rare variant, next generation sequencing, epidemiology

Recent studies show that rare variants play an important role in complex disease etiology. New technologies now allow generating exome sequencing data to detect association between common disease and rare variant mutations. Methods for association studies using common SNPs have low power when being used to to identify rare disease-causing variants due to their low frequency. Most current strategies proposed for detecting rare variants either group or collapse the variants within a region, such as genes or pathways. Although such methods have reasonable power in detecting causal variants, we find that they often have biased control of the Type I error rate. In this paper, we propose a new method to analyze rare variant association data for complex traits.The control of the Type I error rate and the power of the proposed method will be compared to the existing methods for a variety of underlying genetic models.

Secondary Analysis In Gwas

✦Huilin Li, New York University, 650 First Ave Room 547, New York, NY 100016, *huilin.li@nyumc.org*; Mitchell Gail, National Cancer Institute

Key Words: secondary analysis, case-control study, GWAS, maximum likelihood estimation

Case-control genome-wide association studies provide a vast amount of genetic information that may be used to investigate secondary phenotypes. We study the situation in which the secondary phenotype and genetic markers are dichotomous. We first prove that with disease rate is known, the inverse-probability-of-sampling-weighted (IPW) regression method is exactly the maximum likelihood estimation method using the full disease model. Those two methods are the most robust methods in term of guarding the possibility of interaction effect of genetic variants and secondary phenotype on the disease. When there is no interaction effect, the maximum likelihood estimation method with the no interaction assumption is the most efficient method. To strike a balance of the above methods, we proposed an adaptively weighted method that combines the IPW and MLE with reduced disease model. Our adaptively weighted method is always unbiased and has reduced mean square error for estimation with a pre-specified gene and increase the power to discover a new association in a genome-wide study when non-zero interaction is possible. Case-control study with known population totals is also investigated rega

A Wavelet-Based Historical Functional Linear Mixed Model For Examining The Acute Health Effects Of Pollution Exposure

✦Elizabeth Malloy, American University, 4400 Massachusetts Ave. NW, Washington, DC 20016, *malloy@american.edu*; Brent Coull, Biostatistics Department, HSPH; Jeffrey S Morris, The University of Texas MD Anderson Cancer Center; Sara Adar, University of Michigan School of Public Health; Helen Suh, Harvard School of Public Health; Diane Gold, Harvard School of Public Health

Key Words: Bayesian methods, Historical functional linear model, Functional data analysis, Mixed Model, Wavelets

Data measured over time on a grid of discrete values collectively define a functional observation. Studies examining the acute health effects of pollution exposure are frequently interested in the prediction of a functional health outcome from a functional exposure measurement, such as airborne particulate matter or black carbon. We develop a wavelet-based historical functional linear mixed model which allows for the modeling of repeated measures of a continuous functional response and a functional predictor. We employ a novel use of wavelet-packets in a Bayesian setting to regularize the regression coefficient surface and force the historical time constraint, so that future values of the covariate function are not used to predict current or past values of the response function.

Confidence Intervals For The Ratio Of Two Poisson Rates

✦Yonggang Zhao, Pfizer, PA 19403, *yonggangzhao69@yahoo.com*; Qianqiu Li, Johnson & Johnson

Key Words: Confidence interval, Poisson rate, Delta method, Fieller's theorem, Bayesian inference, Bootstrap

This paper describes four methods for constructing a confidence interval for the ratio of two Poisson rates: Delta and Fieller's methods under normal Approximation, Bayesian and Bootstrap methods. These methods are evaluated on simulation data via confidence interval length and coverage probability. Some recommendations are provided under simulation scenarios.

186 Topic Contributed Oral Poster Presentations: Scientific and Public Affairs Committee Competition

Scientific and Public Affairs Advisory Committee

Monday, August 1, 10:30 a.m.–12:20 p.m.

Causal Inference For Cardiology In The Age Of Electronic Health Records

✦David Shilane, Stanford University Department of Health Research and Policy, 94305, *dshilane@stanford.edu*

Key Words: causal inference, observational studies, cardiology, biostatistics, electronic medical records, health research and policy

The advent of electronic health records has greatly improved our ability to assess the quality, safety, and cost of medical treatments. In collaboration with a team of Stanford cardiologists and Kaiser Permanente's Division of Research, I have utilized a variety of electronic databases to provide statistical guidance in medical studies. Kaiser's cardiology records provide unprecedented patient information that integrates clinical, hospital, and pharmacy records with long--term followup. These registries provide larger sample sizes, richer covariate profiles, and greater population representation than randomized clinical trials at significantly reduced cost. However, because the data do not arise from designed experiments, causal inference methods are necessary to account for issues of treatment selection bias. I will present case studies that examine the safety of spironolactone treatments, provide a comparative cost--benefit analysis of imaging tests, and assess the factors

impacting the prescription of, patient adherence to, and effectiveness of medications for ischemic heart disease. Within these studies, I will compare alternative methods for causal inference.

Creating R System For Optimization Of Real-Time Physiologic Parameters: Application Of Loess Regression

✦Joshua Callaway, University of Arkansas for Medical Sciences, 4301 West Markham, Slot 781, Little Rock, AR 72205, *joshcllw@gmail.com*; D. Keith Williams, University of Arkansas for Medical Sciences; Zoran Bursac, University of Arkansas for Medical Sciences; Benjamin Stone, Sigma Human Performance

Key Words: Loess Regression, R, Nonparametric Regression, Nonlinear Regression, VO2 Max

Customized models inform individuals on parameter interpretation during exercise in order to maximize fat contribution to overall caloric expenditure. Due to R's free open-source nature, we have augmented established statistical techniques into a unique system available on demand for parameter optimization. Loess regression serves as an excellent tool for this undertaking due to its flexibility. It takes the best aspects from both linear least squares and nonlinear regression. Loess operates by fitting local models to smaller pieces of the entire dataset, thus adapting to changes in variation throughout the regression. Therefore, we have implemented the R loess command into a function that can plot one or two response variables against the desired predictor variable for individual datasets. In essence, an individual wishing to optimize fat burning can utilize this system in real-time to determine where to maintain intensity during exercise.

Continuous Positive Airway Pressure (Cpap) Adherence In A Sample (N=208) Of Male Military Veterans Diagnosed With Sleep Apnea

✦Matthew R Marler, Independent Consultant, 24931 Rio Verde Drive, Ramona, CA 92065, *matthew.marler@gmail.com*

Key Words: CPAP, Continuous Positive Airway Pressure, Treatment Adherence, Sleep Apnea, Functional Data Analysis

A sample of 208 mostly male military veterans who were diagnosed with sleep apnea were prescribed Continuous Positive Airway Pressure (CPAP), and recorded their adherence with the treatment protocol daily for a term of one year. CPAP adherence is defined as the amount of time spent at the prescribed pressure level, and is measured by the CPAP unit. The data records display diverse features, not all of which have been previously reported in CPAP adherence studies: immediate termination, gradual termination over periods ranging from weeks to months, interruptions after periods ranging from weeks to months followed by resumption of use, alternations of much and little adherence, gradual increase in use over periods of months, high day-to-day variability with near constant mean use over the full recording period. In order to quantify these features and relate them to medical, social and psychological covariate measures, we projected the individual records onto a basis of b-splines with 3 knots at time 0 and interior knots at 30 or 31 day intervals.

Sparse Principal Component Analysis (Spca) Of Wheat Microarray Data Identifies Co-Expressed Genes Differentially Regulated By Cold Acclimation

✦Amrit B Karki, SDSU, 2220 10th Street, #8, Brookings, SD 57006 US, *akarki001@yahoo.com*

Key Words: microarray, SPCA, Fold Change, preprocessing, expression

DNA microarray technology is a powerful tool for high-throughput analysis that has been used for the purpose of monitoring expression levels of thousands of genes simultaneously and identifying those genes that are differentially expressed. The high dimensionality of microarray data, the expression of thousands of genes in a much smaller number of samples, presents challenges that affect the validity of analytical results. A main issue in microarray transcription profiling is data mining and analysis. Statistical methods are vital for these scientific endeavors. Here three methods are used to compare data obtained from a Transcriptome analysis of cold acclimation effects on two lines of winter wheat varying in freeze survival. The line with 75 % survival was designated FR and the line with 30 % freeze survival is designated FS. The three methods consisted of fold change (FC), moderated t-test and sparse principal component analysis (SPCA). After pre-processing of the microarray data, total number of genes left after pre-processing and presence call selection was 25,770 expressed in FR and 26,264 expressed in FS. These genes were further divided into those expressed in FR only (2

Optimal Variable Weighting In K-Means Clustering

✦Shaonan Zhang, State University of New York at Stony Brook, Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11790, *zshaonan@gmail.com*; Jiaqiao Hu, State University of New York at Stony Brook; Wei Zhu, State University of New York at Stony Brook

Key Words: K-means clustering, variable weights, optimization, principal component analysis (PCA)

K-means clustering method is a widely adopted classic clustering algorithm. Weighted K-means clustering is an extension of the K-means clustering in which a set of nonnegative weights are assigned to all the variables. In this paper, we aim to derive the optimal variable weights for weighted k-means clustering in order to obtain more meaningful and interpretable clusters. We further optimized the weighted k-means clustering method (MH Huh, YB Lim 2009) by introducing a new algorithm to obtain global-optimal guaranteed variable weights based on the Karush-Kuhn-Tucker conditions. Here we first present the related theoretical formulation and derivation of the optimal weights. Then we provide an iteration-based computing algorithm to calculate such optimal weights. Numerical examples on both simulated and real data are provided to illustrate our method. It is shown that our method outperforms the original proposed method in terms of classification accuracy and computation efficiency. Finally, a modified solution based on the principal component analysis is proposed to further improve the computational efficiency of K-means clustering for data set with a large number of variables.

Estimating The Effect Of Dust Events On Daily Hospitalizations For Asthma While Adjusting For Hourly Levels Of Air Pollutants With The Historical Functional Linear Model

✦Priyangi Kanchana Bulathsinhala, University of Texas at El Paso, 500 West University Avenue, University of Texas at El Paso, El Paso, TX 79968, *pkbulathsinhala@miners.utep.edu*; Joan Staniswalis, University of Texas at El Paso; Sara Grineski, University of Texas at El Paso

Key Words: Nonparametric functional linear model, Case-crossover design, Conditional logistic regression

El Paso, Texas is known as one of the dust "hotspots" in North America. We explore the effect of dust storms on asthma admissions in El Paso, Texas between 2000 and 2005. Conditional logistic regression with a case-crossover design was used to estimate the probability of hospitalization during dust events while controlling for pollutants with hourly monitor measurements, and weather. The functional linear model is used to incorporate the hourly pollutant measures into the regression model with a continuous lag, as an alternative to a distributed lag model based on daily averages. The nonparametric functional linear model in the conditional logistic regression framework is fit by first preprocessing the data, then applying the COXPH function in the R-package for survival analysis with a slight modification. We use the ridge trace to guide the choice of the smoothing parameter in nonparametric functional linear model. This is described and preliminary findings are reported.

Investigating The Potential Relationship Between Change Of Traffic Volume On Toll Roads And The Change Of Gas Price

✦Ranye Sun, Texas A&M University, Department of Statistics, 3143 TAMU, College Station, TX 77843, *rsun@stat.tamu.edu*

Key Words: Autoregressive Distributed Lag (ADL) model, Autoregressive Error, Gibbs sampling, Metropolis algorithm

Travelers' response to changes in the cost of travel provides key data to help predict future travel behavior. The objective of this paper is to investigate the potential relationship between change of traffic volume on toll roads and gas price change and other economic factors. Two models were used in the project, Autoregressive Error (AE) model and Autoregressive Distributed Lag (ADL) model. In this paper, Bayesian Analysis was used to generate data from posterior distributions with non-informative prior and informative prior. Gibbs sampling with Metropolis algorithm was applied and posterior modes, analysis for single parameters and bivariate parameters were found as the inference. Statistical analysis and economic analysis of the model selection, model conditions, model estimation and prediction are given.

On Misspecification Of Transition Models With Garch Error Term: The Monte Carlo Evidence

✦OLANREWAJU I SHITTU, DEPARTMENT OF STATISTICS, UNIVERSITY OF IBADAN, IBADAN, IBADAN, International 23402 NIGERIA, *oi.shittu@mail.ui.edu.ng*; OLAOLUWA S YAYA, DEPARTMENT OF STATISTICS, UNIVERSITY OF IBADAN

Key Words: Specification test, Smooth transition autoregression, Serial correlation, Taylor's series approximation

Smooth transition models have gained popularity in modelling economic and financial series due it ability to capture the non linearity in the data sets. However, nonlinear time series with serial correlation in its error term which sometimes leads to misspecification of appropriate models for important financial or economic data has not been considered. This paper then considers the effect of a first order serial correlation of the residuals on the nonlinear time series model by specifying a variant of smooth transition model ESTAR using the Monte Carlo simulation methods with a view to examining correct specification with Escribano and Jord· (EJP) specification procedure under various levels of nonlinearity and varying degree of standard deviation of the data. It is established that the power of misspecification of non linear model is a function of serial correlation of the residuals, the sample size, degree of nonlinearity and the standard deviation of the series. Correct model is specified at moderate values of standard deviation and serial correlations with great cautions. The results will then results will serve as guide in empirical econometric and time series modelling.

Applications Of The K-Th Year Moving Average Approach To Time Series Forecasting Of Breast Cancer Mortality Rates

✦Michael Kotarinos, University of South Florida, 1608 Allen's Ridge Drive North, Palm Harbor, FL 34683, *mkotarin@mail.usf.edu*; Chris Tsokos, University of South Florida

Key Words: ARIMA, Time Series Analysis, Breast Cancer, Moving Average

The object of the present study is to apply a newly developed time series forecasting technique to yearly breast cancer mortality rates to develop a statistical model that characterizes breast cancer mortality over time. Using SEER cancer data, we develop a K-th Year ARIMA model for breast cancer mortality that reveals how breast cancer develops over time. By analyzing this time series, we determine the effectiveness of strategic initiatives to combat breast cancer and predict future trends in breast cancer mortality. Using residual analysis, the quality and effectiveness of the models were evaluated in regards to their predictive power and applicability in comparison to standard regression models.

Cross-Sectional And Longitudinal Joint Modeling Of Repeated Measures Of Quasi-Continuous Patient-Reported Outcome And Binary Response Data

✦Kelly H. Zou, Pfizer Inc, 235 East 42nd Street, Mail Stop 219/08/02, New York, NY 10017, *Kelly.Zou@pfizer.com*; Jeno P. Martin, Pfizer Inc; Richard J. Willke, Pfizer Inc

Key Words: Randomized Controlled Trial, Patient-Reported Outcome, Receiver-Operating Characteristic Curve, Trajectory Analysis, Mixed-Effects Model, Probit Regression

It is challenging to compare the treatment effects based on quasi-continuous patient reported outcome (PRO) scores and a transient binary response indicator on the same subjects in randomized control trial with repeated-measurements. We propose several methods to jointly model the PRO domain score and binary response. Cross-sectionally, two-sample tests are performed on the PRO domain scores between

responses within each treatment. A receiver operating characteristic curve analysis, with the response as a gold standard, is used to estimate an optimal threshold along the PRO measurement scale using the Youden index. Longitudinally, cumulative distribution functions of actual and percent change from baseline of a PRO are visual displays of the entire responses. Trajectory analysis of the time-course of the PRO is performed for each treatment. A mixed-effects model of PRO domain scores is conducted to adjust for baseline covariates. Finally, bivariate probit regression and generalized linear mixed modeling of the PRO and response assesses the effect of treatment and additional baseline covariates. These methods are illustrated on a PRO instrument and a time-varying binary response.

Adaptive Intra Patient Escalation Design For Phase I Trials In Oncology

✦Laura Levette Fernandes, University of Michigan, 327 S Division Apt b, ann arbor, MI 48104, *flaura@umich.edu*; Jeremy Michael George Taylor, University of Michigan

Key Words: adaptive design, dose ranging, Markov model, Bayesian analysis., oncology, stochastic

Phase I Dose escalation studies in oncology focus on finding the maximum tolerable dose (MTD) that gives an acceptable levels of toxicity. Often times the therapy is given in multiple cycles with breaks between the cycles. Usually the dose assignment for a patient does not change from one cycle to the next and the toxicity is determined from the first cycle only or from the whole regimen. An alternative design is to allow intra-patient variation in dose and to assess the toxicity separately in each cycle. In this scheme the next dose assigned to a patient would be escalated or deescalated based on past responses from this and other patients and driven by a statistical model. We propose a stochastic model for analyzing such data. The model is a 3 parameter Markov model in which the probability of toxicity at cycle j given no prior toxicity is a function of the cumulative dose, the current dose and the maximum of the previous doses administered to the patient. The data is then analyzed in a Bayesian way using Winbugs. Correct modeling of the data and proper estimation of the parameters enables projecting the outcomes of the trial when different doses are assigned at the next stage.

Keeping Pace With Changes In Society - A Glance At Selected Questions Tested For Canada'S 2011 Census And National Household Survey

✦Fil McLeod, Statistics Canada, 170 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6 Canada, *fil.mcleod@statcan.gc.ca*

Key Words: social, census, qualitative testing, statistical testing

This poster presentation will look at selected questions tested for the 2011 Census and the new National Household Survey that replaces the previous long-form census. The presentation will include examples such as evolution of family and commuting to work. It will answer the following questions: What triggered the content change? What criteria were considered for implementing the change? How did the question fare in qualitative and statistical testing? What outcome can be expected?

Construction Of A Coincident Index By Means Of Dynamic Common Factors

✦WILMER OSVALDO MARTINEZ RIVERA, CENTRAL BANK, CARRERA 7 # 14 - 78, BOGOT¡, International 4-74 COLOMBIA, *womartinezr@gmail.com*

Key Words: coincident index, common factors, coincident profile

The model proposed in PeÒa & Pocela (2006) allows to extract the dynamic part that is common to a set of time series, which may be non-stationary, but the question is which of them can explain a economic state, finances or the climate; to this end we develop a methodology to establish which of those factors better tracks a reference series.

Medical Tests For The Diagnosis Of Myasthenia Gravis: A Statistical Analysis Of Jitter On Single Fiber Electromyography.

✦Norberto Pantoja, US Food and Drug Administration, MD , *norberto.pantoja-galicia@fda.hhs.gov*; Rebecca A. Betensky, Harvard School of Public Health, Harvard University; Pushpa Narayanaswami, Harvard Medical School; Seward Rutkove, Harvard Medical School

Key Words: Early stopping, futility analysis, conditional power design, medical diagnostic tests, myasthenia gravis

Single Fiber Electromyography (SFEMG) is an electromyographic technique that identifies and records action potentials from individual muscle fibers. The major clinical application of SFEMG has been in the diagnosis of Myasthenia Gravis. During SFEMG at least 20 muscle fiber action potential pairs are recorded from a given muscle. An abnormal study can be declared if at least 10% of potential pairs have jitter that exceeds the upper limit of normal for paired jitter in that muscle. SFEMG is a time consuming procedure associated with discomfort. If it were possible to predict that the test was likely to be normal after a certain number of muscle fiber pairs had been analyzed, the test could be terminated earlier without the extra time and discomfort associated with collecting 20 fiber pairs. There would be little motivation to continue the procedure if the probability of declaring an abnormal test by the end of the 20 pairs given the interim data is low. We propose a statistical analysis that suggests early stopping of the SFEMG procedure if the conditional probability of declaring an abnormal test by the end of the procedure given the interim data is sufficiently low

Small Area Estimation Of County Level Cash Rent Rates

Will Cecere, National Agricultural Statistics Service; ✦Emily Berg, National Agricultural Statistics Service, 22031, *Emily_Berg@nass.usda.gov*; Malay Ghosh, University of Florida

Key Words: small area estimation, mixed model,, mean square error

The National Agricultural Statistics Service (NASS) collects cash rent data through an annual Cash Rents Survey and produces estimates of cash rent rates by several agri-land types at the county level. A cash rent is land that is rented on a per acre bases for cash only. The county level cash rent estimates are used to set payment rates for the Conservation Reserve Program. The cash rent estimates are also useful to farmers in determining rental agreements. Small area estimation techniques are proposed as a way to improve estimates of county level cash rent rates.

Estimates for 2010 are obtained using data from the 2009 Cash Rents Survey and external sources of information as auxiliary variables. Results are presented and mean square error estimation is discussed.

Testing For The Shape Of A 'Polyhedral' Cellular Inclusion

✦Sukantadev Bag, University College Cork, Department of Statistics, University College Cork, Western Gateway Building, Western Road, Cork, International 00 Ireland, *sukantadev.bag@gmail.com*; Kingshuk Roy Choudhury, Statistics Department; Mingzhi Liang, University College Cork; Michael Prentice, University College Cork

Key Words: Cryo-electron tomography, shape analysis, polyhedron, principal component analysis

We consider the problem of identifying the shape of a type of cellular inclusion, called metabolosome, newly identified in E. coli cells from 3d reconstructions obtained via cryo-electron tomography. Due to limited angle tomography, the reconstructions are of uneven resolution, with a missing wedge at the top and bottom of the reconstruction. Most reconstructed slices appear to have convex polygonal shapes, suggesting a convex polyhedral shape for the 3d object. Naturally occurring objects such as viruses have previously been noted to possess symmetric polyhedral shapes. However, principal component analysis of a collection of metabolosomes reveals variation and a significant departure from symmetry. To test the hypothesis that the observed objects may be deformed Archimedean solids, we use the edge profile statistic, which is a vector of the number of edges of polygons in successive slices of the object. This statistic is invariant under shear transformations and uniquely identifies Archimedean solids. Resulting tests can be applied under missing data. Our tests reveal significant departures from all Archimedean solids.

187 Contributed Poster Presentations: ENAR

ENAR

Monday, August 1, 10:30 a.m.–12:20 p.m.

Publication Bias In Meta-Analysis

✦Min Chen, ExxonMobil Biomedical Sciences, Inc, , *min.chen@exxonmobil.com*

Key Words: Meta Analysis, Publication Bias, Observational studies, funnel plot, Trim and Fill, Cumulative meta-analysis

Meta-analysis is the statistical synthesis of results of several studies that address a set of related research hypotheses. The synthesis will be meaningful only if the studies have been obtained systematically. However, publication bias is an issue for meta-analysis. Studies with significant results are more likely to be published than those with non-significant results, and published studies are more likely to be included in a meta-analysis than unpublished studies. In this presentation I will address publication bias in meta-analysis for observational studies using various statistical methods including the funnel plot, Rosenthal's Fail-safe, Orwin's Fail-safe N, Duval and Tweedie's Trim and Fill, and the cumulative meta-analysis.

The Practical Effect Of Batch On Genomic Prediction

✦Hilary Parker, Johns Hopkins School of Public Health, 615 N Wolfe St E3035, Department of Biostatistics, Baltimore, MD 21205 US, *hiparker@jhsph.edu*; Jeff Leek, Johns Hopkins School of Public Health; Rafael Irizarry, Johns Hopkins School of Public Health

Key Words: Genomics, Batch Effects, Microarrays, High-throughput, Prediction, Bioinformatics

Measurements from microarray and other high-throughput technologies are susceptible to a number of non-biological variables such as temperature and reagent lots. It has been shown that these artefacts, collectively called batch effects, can severely alter the outcome of any differential expression analysis, resulting in misleading biological conclusions. Here we examine the impact of batch effects on predictors built from genomic technologies - specifically gene expression microarrays. We compare single microarray (fRMA) and multiple microarray (RMA, MASS5) preprocessing methods and both rank-based (top-scoring pairs) and continuous (PAM) predictors. We show that in general, prediction is made more difficult by batch effects. We also show that when there is perfect confounding of batch and the outcome being predicted, then accuracy is substantially reduced. In an effort to mitigate this effect, we determine which probes from commonly used Affymetrix arrays are most susceptible to batch and investigate their properties. Down-weighting these "batch-affected" probes may lead to increased predictive accuracy when building gene expression based predictors.

The Validity And Power Of Extreme Sampling Schemes For Mediation Analysis

✦Robert Makowsky, University of Alabama at Birmingham, RPHB 327, 1530 3rd AVE S, Birmingham, AL 35294 US, *rmakowsky@ms.soph.uab.edu*; Mark Beasley, University of Alabama at Birmingham; Gary L. Gadbury, Kansas State University; Jeffrey M. Albert, Case Western Reserve University ; David B. Allison, University of Alabama at Birmingham

Key Words: Mediation, Sampling, Missing data

In some cases, the costs of measuring the putative mediator (Z) or the outcome (Y) can be large. Extreme sampling procedures may reduce study costs by increasing power per subject measured on the more expensive variable. Therefore, we explored whether sampling schemes will produce reliable parameter estimates and increase power for fixed study cost or, complementarily, decrease cost for fixed study power. Three sampling strategies were considered; 1) sampling from above and below the 75th and 25th percentiles, respectively, of Z ('quartile-sampling'), 2) quartile-sampling from Z conditional on the study arm, and 3) removal of half of the specimens based on ordering of Z. Simulation results showed that all sampling schemes produced biased correlation coefficients when analyses were based on specimens with complete data, leading to inflated Type 1 Error rates. When maximum likelihood was used to handle missing data, bias in point estimates was eliminated. For all sampling approaches, an increase in the variance of most correlation estimates was observed, and this may account for Type 1 Error rate inflation in some situations, despite lack of bias in the correlation coefficients.

A Comparative Study Of Power Of The Association Tests For Linkage Disequilibrium

✦Sayan Dasgupta, University of North Carolina at Chapel Hill, 416 West Cameron Ave, Apt #2, Chapel Hill, NC 27516 US, *sdg.roopkund@gmail.com*

Key Words: TDT, HHRR, Hardy Weinberg Equilibrium

The transmission disequilibrium test (TDT) was proposed as a family-based association test for the presence of genetic linkage between a genetic marker and a trait. The idea was generated from the proposition that data be gathered in trios of parents plus an affected child. An earlier method (HHRR) used these trios to model the transmitted vs. nontransmitted alleles as units of observation in a haplotype-based haplotype relative risk approach. Considering a recessive disease model in this study, we will use simulation and analytic results to show that: i) The HHRR method is more powerful to detect true association than TDT under the classic situation with 2 alleles and both parents fully observed in case-parent trios assuming that Hardy Weinberg Equilibrium holds. ii) When parental mating choice is not random, so that heterozygous individuals are more likely than expected under HWE in the parental population, the HHRR statistic based on the chi-square(1) null approximation can inflate the Type I error. iii) And lastly we propose an empirical based method that draws on the strengths of both HHRR and TDT, and is more powerful than TDT under modest departures from HWE in parental.

Comparison Of Association Methods For Genetic Association Studies

✦Olivia Bagley, North Carolina State University, 27695-7566, *okbagley@ncsu.edu*; Allison Huber, Meredith College; Rachel Beckner, Meredith College; Wesley Stewart, North Carolina State University; David Reif, North Carolina State University; Alison Motsinger-Reif, North Carolina State University

Key Words: statistical genetics, Multifactor Dimensionality Reduction, LASSO regression, logistic regression

Detecting variants that predict complex disease is a major goal of human genetics, but is a difficult challenge due to complex underlying etiologies. There are several potential models for the genetic etiology of complex traits, but sources of noise in the data set such as heterogeneity, epistasis, and missing data present challenges for statistical association methods. The effect of these different error-types is largely unknown for data-mining approaches. The impact of each factor needs to be understood to properly apply and modify data mining approaches in human genetics. We simulated data representing a wide range of genetic models to assess the ability of commonly used association analysis approaches to correctly identify the underlying genetic model in the presence of different types of error. Specifically, we compared the power of traditional and stepwise logistic regression, Multifactor Dimensionality Reduction, and LASSO regression and show the relative performance of each method in the presence of these types of noise. This material is based upon work supported by the National Science Foundation under the NSF-CSUMS project DMS-0703392 (PI: Sujit Ghosh).

Effectiveness Of Grammatical Evolution Decision Trees In Identifying Disease Causing Polymorphisms

✦Kristopher Hoover, North Carolina State University, , *kmhoover@ncsu.edu*; Rachael Marceau, North Carolina State University; Tyndall Harris, North Carolina State University; David Reif, North Carolina State University; Alison Motsinger-Reif, North Carolina State University

Key Words: evolutionary computation, statistical genetics, gene-gene interactions, decision trees, variable selection, genetics

The identification of complex genetic models including gene-gene interactions that predict common disease is an important goal of human genetics. Detecting such complex models in high-dimensional data creates an analytical challenge, requiring methods to perform both statistical estimation and variable selection. Recently, a Grammatical Evolution Decision Tree (GEDT) approach has been proposed to detect such complex models. Decision trees are easily interpretable; however their power is limited in identifying strict interactions due to their usual hierarchical model building approach. Grammatical evolution, a type of evolutionary computation, is used to avoid this problem by evolving decision trees to detect interactive models. In the current study, we present the results of parameter sweep optimization for GEDT such as generations, mutation rate, etc. Additionally, we show power comparisons to similar methods such as neural networks and logistic regression approaches. This material is based upon work supported by the National Science Foundation under the NSF-CSUMS project DMS-0703392 (PI: Sujit Ghosh).

Using Haplotype Blocks For Detecting Interactions With Multifactor Dimensionality Reduction

✦James Kniffen, North Carolina State University, , *jkniffenjr@gmail.com*; Nicole Mack, North Carolina State University; Alison Motsinger-Reif, North Carolina State University; David Reif, North Carolina State University

Key Words: statistical genetics, haplotypes, Multifactor Dimensionality Reduction, gene-gene interactions, linkage disequilibrium

All genetic association analysis relies on linkage disequilibrium (LD), which describes the nonrandom assortment of genomic variants which is quantified as correlation between variants. This redundancy allows large sets of single-variant genetic data to be collapsed into haplotype blocks, which identify all correlated variant sites in a particular genomic region. Using haplotypes instead of single-variants can increase the power of gene-mapping studies due to denser data and reduction in number of tests. This is an important advantage as the scale of genetic studies creates challenges in variable selection, especially when searching for complex models including gene-gene interactions. In the current study we propose a haplotype collapsing approach to reduce the number of input variables for a commonly used approach for detecting interactions: Multifactor Dimensionality Reduction. We investigate this approach with simulated data to evaluate its performance and power to detect complex predictive models. The research is based upon work supported by the National Science Foundation under CSUMS grant #DMS-0703392 (PI: Sujit Ghosh).

Profiling Center-Specific Long Term Kidney Transplant Outcomes

✦Kevin He, University of Michigan, Ann Arbor, MI 48103, *kevinhe@umich.edu*; Douglas E. Schaubel, University of Michigan

Key Words: Center effect, Direct standardization, Laplace approximation, Mean survival time, Piece-wise exponential, Log-normal frailty model

Current measures of transplant center specific performance focus on short term outcomes, despite the fact that most graft losses occur in the long term. In this paper, we propose a method which combines a log-normal frailty model and piece-wise exponential baseline rate to compare the mean survival time across centers. The challenging part is that multi-center studies tend to generate heterogeneous samples (unequal covariate effects across centers). Laplace's approximation for integration is applied for maximum likelihood estimations. Direct and indirect standardization methods are applied to estimate the mean survival time. The proposed methods allow for valid estimation of mean survival time as oppose to the restricted mean lifetime and, within this context, robust profiling of long-term center specific outcomes. Asymptotic properties of proposed estimators are discussed. Finite-sample properties are examined through an extensive simulation study. The strengths and weaknesses of direct and indirect standardization are compared. The methods developed are applied to national kidney transplant data. The best (and worst) 10% of transplant centers in the Unites States are identified.

Comparison Of Rppa Data Normalization Methods

✦Wenbin Liu, UT MD Anderson Cancer Center, , *wliu@mdanderson.org*

Key Words: RPPA, normalization, protein arrays

The reverse phase protein array (RPPA) is a technology that measures protein expression in hundreds to thousands of samples at once. The EC50 of each dilution series of the sample lysate is usually estimated by fitting a nonlinear model. Our goal in this study is to compare some of the existing and potentially methods and some of the methods that we have proposed, including global median centering the samples (MCS), median centering the samples and then median centering the slides (MCSS1) and vice versa (MCSS2), median polish (MP), centering around total protein amount (measured by cloidal gold), variable slopes (VS) and z-score transforming each slide and then median centering each sample (ZS). Cell line data and simulated data were used. After normalizing the data by different methods, we compared the correlation between the replicate slides and the correlation between the replicate samples, the sample median ranks and the sample rank MADs, and biological validations using the cell line data and drug treatment information. The MCS and VS methods outperformed the other methods in precision and accuracy.

188 Contributed Oral Poster Presentations: Section for Statistical Programmers and Analysts

Section for Statistical Programmers and Analysts
Monday, August 1, 10:30 a.m.–12:20 p.m.

Using Proc Template To Convert Sas Data To Define.Xml

Matthew Wiedel, Celerion; ✦Lucius A Reinbolt, Celerion, 621 Rose St., Lincoln, NE 68502, *lucius.reinbolt@celerion.com*; Aleksandra Stein, Celerion; Vanessa Huang, Celerion; Steven Kirby, Celerion; Nancy Wang, Associate Director of Biostatistics

Key Words: PROC TEMPLATE, Define.xml, Metadata

Well formed data contains all information necessary to understand study results and conclusions, but that information is not accessible to a general audience. For data to gracefully speak to end users, further information is needed. A define.xml is the additional resource the reader needs to fully appreciate the data. Based on FDA guidance, a define.xml document is now the default standard for use with CDISC SDTM data. A two-step data-driven process is used to create a SAS define.xml document. The first step is the metadata creation which produces 6 SAS define data sets. A data mining approach is utilized to describe the many levels of information within the data. The next step is the integration of SAS define data into an xml formatted document utilizing a robust SAS PROC TEMPLATE approach. Sorted metadata is restructured and transformed into a group of tagsets that comprise the define document. The benefit of this approach is the concise and powerful application of PROC TEMPLATE. It presents the data through a series of define events triggered by the levels of metadata. The proc template code is simple to alter and run, allowing more time to focus on the data itself.

Displaying Small Sample Size Datasets Effectively

Thomas E Bradstreet, Merck Co; Shuping Zhang, Merck Co; ✦Xingshu Zhu, Merck Co, 351 N. Sumneytown Pike, North Wales, PA 19454, *xingshu_zhu@merck.com*

Key Words: sample size, dot plot

Visualizing individual data, empirical distributions, and summary statistics for small to moderate sized samples is critical for effective statistical analysis and inference in early phases of drug research. We present a customized SASÆ macro, %HangingDotPlot, which displays dot charts with either parametric or nonparametric descriptive statistics and confidence intervals, for each of multiple bins of data displayed simultaneously on the same page. The flexibility and utility of the macro, including back-to-back dot chart presentations, is demonstrated with visual analogue scale data from human trials and swim maze testing data in rats.

A Single Imputation Sas Macro

✦Xingshu Zhu, Merck Co, 351 N. Sumneytown Pike, North Wales, PA 19454, *xingshu_zhu@merck.com*; Shuping Zhang, Merck Co

Key Words: missing data, imputation

The problem of missing data arises in many clinical trials, and practical research fields. Imputation is often required to replace the missing value of a variable in a dataset such the "completed" data set can be used in subsequent analyses of the data. When working with an extensive dataset containing millions of records, however, it would be highly impractical, or maybe even impossible, to use the multiple imputation method because of the overwhelmingly large number of datasets that would have to be created. In this paper, we introduce a simple SAS macro that allows the user to create a "complete" SAS dataset through single imputation by selecting different statistical methods.

There Is Such A Thing As Too Much Data

✦Ying Su, Merck Research Laboratories, Merck & Co., Inc., Upper Gwynedd, PA 19454 United States, *ying_su@merck.com*

Key Words: SAS, Pharmacoepidemiology, Nested Case-Control Study, Random Sampling

Pharmacoepidemiology studies increasingly use healthcare databases, which involve massive amounts of data. Sometimes conventional programming methods are unable to handle the data volume, or even overwhelm the computing resources. An example of selecting controls for a nested matched case-control study is used to illustrate this challenge. Different attempted approaches in SAS programming are discussed, and a solution is developed to overcome the limitation of computing space and speed. The solution presented also addresses the programming efficiency concerns, and ensures the statistical randomness in sampling. Lessons learned from this example will provide guidance for other studies with similar characteristics.

Shrinkage And Absolute Penalty Estimation In Linear Models

✦SM Enayetur Raheem, University of Windsor, 401 Sunset Ave, Windsor, ON N9B3P4 Canada, *raheem@gmail.com*; Ejaz Syed Ahmed, University of Windsor

Key Words: Shrinkage estimation, Absolute Penalty Estimation, Linear Regression, James-Stein Estimator

We study the James-Stein-type shrinkage estimation of regression parameters in a linear regression setup. Shrinking of the parameters towards both the null vector and a sub-vector is considered. We conduct Monte Carlo simulation to compare, under quadratic risk criterion, an absolute penalty estimator and James-Stein-type shrinkage estimator. Analytic results show that shrinkage estimators demonstrate asymptotically superior risk performance relative to the classical estimator. Our Monte Carlo study reveals the superiority of shrinkage estimators over absolute penalty estimators in the presence of a relatively large number of nuisance covariates. Application to a real life data set will be provided.

A Sas Macro To Create An Application To Calculate Sample Size For Evaluating Mediation Analysis: The Linear Model Case

✦Rajendra Kadel, University of South Florida, , *rkadel@mail.usf.edu*

Key Words: SAS, Mediation Analysis, Linear regression, sample size, MACRO Program, application

Mediation analysis is used to review the comparative change in the amount of strength of association of the primary predictor with the outcome after adjustment for the mediator. Mediation models are very widely used in social and biomedical sciences. Before conducting mediation studies, researchers want to know the sample size required for achieving the adequate power when testing for mediation. To the author's knowledge, there is no any SASÆ procedure that produces sample size calculation for mediation analysis. The author presents a macro to create an application to calculate sample size for mediation analysis. It implements the methods to calculate sample sizes for the linear regression models. Very basic understanding of SAS is enough to use this macro application. %WINDOW and %DISPLAY macro statements along with a SAS macro are used to create this application. However, SAS Macro programming skill is not expected. When the macro is invoked, a series of windows will pop-up asking user to input required information. Output will be presented in SAS output window as well as in Microsoft word document using ODS RTF statement.

189 Contributed Oral Poster Presentations: Section on Statistics in Epidemiology

Section on Statistics in Epidemiology

Monday, August 1, 10:30 a.m.–12:20 p.m.

Applying Competing Risks Analysis Methods to a Population-Based Study

✦Hyun Ja Lim, University of Saskatchewan, 107 Wiggins Road, Saskatoon, SK S7N5E5 Canada, *hyun.lim@usask.ca*; Xu Zhang, Georgia State University; Roland Dyck , University of Saskatchewan ; Nathaniel Osgood , University of Saskatchewan

Key Words: Competing Risks Analysis, cause-specific hazards model, subdistribution hazards model, Lunn-McNeil model, Diabetes Study

In medical research, each patient can experience one of several different types of events over the follow-up period. Survival times are subject to competing risks if the occurrence of one event prevents the other event types from occurring. The cause-specific hazards and the cumulative incidence functions (CIF) are the most appropriate approaches to analyse the competing risks data. In this study, we compare three models: the cause-specific hazards model, the subdistribution hazards model, and the Lunn-McNeil unstratified model. A population based diabetes study is used to demonstrate the three methods. Our analyses show that the three methods yield different results with regards to the effects of the covariates. The CIF of the cause-specific hazards model reveals a higher CIF curve than the subdistribution hazards model while the CIF of the unstratified Lunn-McNeil model is the lowest. For a dominant risk, one can either use the cause-specific hazards model or the subdistribution hazards model. For a minor risk, the cause-specific hazards model is more appropriate. The Lunn-McNeil method is applicable only when the baseline hazards for the different risk types are proportional.

Changes In Age-Specific Death Rates From Six Leading Causes Of Death In Fifty Us States And The District Of Columbia, 1970-2004

✦Rong Wei, National Center for Health Statistics, CDC, USA, 14003 Marleigh Lane, Cliftion, VA 20124, *rrw5@cdc.gov*; Melonie Heron, National Center for Health Statistics, CDC, USA; Jiraphan Suntornchost, National Center for Health Statistics; Meena Khare, National Center for Health Statistics, CDC, USA

Key Words: US mortality, death rates, leading cause of death, locally weighted regression

The National Center for Health Statistics has well-documented US death data collected through the National Vital Statistics System since 1968. Based on the data from 1970 to 2004, this study demonstrates changes in death rates from six leading causes of death: heart disease, cancer, stroke, accidents, chronic obstructive pulmonary disease and diabetes over 35 years in 50 states and the District of Columbia. Death rates are computed into single years of age, so differences in change over time can be explored across age-, cause of death- and state-specific groups. A locally weighted regression is applied to smooth the rates in two dimensions: age and year of death. Death rates from each leading cause are also compared with ones from all causes combined. This study shows that ranks of leading causes of death vary across ages and time trends of death rates vary across states. Interestingly, while most states follow similar patterns over time, some states have unique trends over time. Studying changes in age-, cause of death- and state-specific death rates dissects the underlying causes of overall mortalities, therefore helps us understand the health variations in US populations.

Estimation Of Dosage And Duration On Drug In Electronic Medical Records

✦Qing Liu, Pfizer Inc, 500 Arcola Rd, Collegeville, PA 19426, *kathy.liu@pfizer.com*; Xiaofeng Zhou, Pfizer Inc; Bing Cai, Pfizer Inc

Key Words: Pharmacoepidemiology, drug utilization study, electronic medical records, average daily dosage

Estimating average daily dosage and duration on the drug is an important element of many Pharmacoepidemiology studies. However, there are many challenges in such estimation in electronic medical records. The challenges we had with analyzing an EMR are: 1) missing data and 2) some dosage information exist solely in string format (e.g., twice a day) and there is no universal algorithm to readily convert such string values into numeric values. In our study, we first focused on a single drug and calculated the proportion of the missing data, and then selected a method to impute missing values after comparing different imputing methods (impute by mean, impute by median, impute based on different subgroups, etc.), and finally tested the efficiency of converting string dosage information into numeric values by using a dosage mapping dictionary and improved this method so that it would be more efficient to be used in future studies. Sensitivity analysis of these methods/algorithm as well as the generalizability of the methods/algorithm to other products are also examined.

A Critical Evaluation Of The Clinical Utility Of Genotype Risk Scores

✦WONSUK YOO, University of Tennessee Health Science Center, 66 N Pauline St, Suite 633, Memphis, TN 38163, *wyoo1@uthsc.edu*; Brian Ference, Wayne State University

Key Words: SNPs, genotype risk scores, log-additive model, linear-additive model, multiplicative combined effect

Substantial uncertainty exists as to whether combining multiple-disease associated single nucleotide polymorphisms (SNPs) into a genotype risk score (GRS) can provide clinically useful information. We critically evaluated the ability of a simple count GRS to predict the risk of a dichotomous outcome, under both a multiplicative (log additive) and linear additive model of combined effects. We also evaluated the ability of a count GRS to predict the level of a continuous outcome variable. We then compared the results of these simulations with the observed results of published GRS measured within multiple epidemiologic cohort studies. A simple count GRS can provide clinically useful information to predict the level of a continuous variable under a variety of circumstances. By contrast, a count GRS is unlikely to be clinically useful for predicting the risk of a dichotomous outcome. Alternative methods for constructing GRS that attempt to identify and include SNPs whose combined effect is at least multiplicative are needed to provide clinically useful information about the genetic risk of dichotomous outcomes.

Estimating The Effect Of Cluster-Level Adherence On An Individual Binary Outcome With A Complex Sampling Design

✦Zhulin He, University of Florida, Department of Biostatistics, Gainesville, FL 32610, *zhulinhe@ufl.edu*; Babette Brumback, University of Florida; Richard Rheingans, University of Florida; Matthew Freeman, Emory University; Leslie Greene, Emory University; Leslie Greene, Emory University; Leslie Greene, Emory University

Key Words: structural nested models, complex sampling design, instrumental variables, estimating equations, unmeasured confounding, global health

We wish to estimate the effect of school-level adherence on individual absenteeism in the context of a school-based water, sanitation, and hygiene intervention in Western Kenya. Schools within strata were disproportionately sampled and randomized to one of three interventions. Next, students within schools were disproportionately sampled and measured for outcomes such as absenteeism. We use double inverse-probability weighting to adjust for the disproportionate sampling and the association of individual-level confounders with randomization. We develop and apply methods based on structural nested models to estimate effects of adherence assessed in terms of relative risks, using school-level randomization as an instrumental variable and using the double weights to adjust for complex sampling and individual-level confounding.

Likelihood-Based Inference For Antedependence (Markov) Models For Categorical Longitudinal Data

✦Yunlong Xie, Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, *yunlong-xie@uiowa.edu*; Dale Zimmerman, University of Iowa

Key Words: Categorical, Longitudinal, Antedependence (Markov) Models, Likelihood-Based Inference

Abstract: Antedependence (AD) of order p, also known as the Markov property of order p, is a property of index-ordered random variables in which each variable, given at least p immediately preceding variables, is independent of all further preceding variables. Zimmerman and Nunez-Anton (2010) present statistical methodology for fitting and performing inference for AD models for continuous (primarily normal) longitudinal data. But analogous AD-model methodology for binary longitudinal data has not yet been developed. In this research, we present methods for determining the order of antedependence of binary longitudinal data. Specifically, we derive the likelihood ratio test, score test, and Wald test for pth-order antedependence against the unstructured (saturated) multinomial model. Simulation studies show that the likelihood ratio test performs better than the others for samples of small and moderate size. We extend the tests for use in testing for pth-order antedependence against qth-order antedependence, where q > p, and we develop a penalized likelihood method for determining variable order antedependence structure. The methods are illustrated using two data sets.

Statistical Modeling The Disappearance Of Helicobacter Pylori And Its Association With Gastric Cancer And Childhood Asthma

✦Wei Wei, University of Minnesota Rochester, Rochester, MN 55904, *wwei@r.umn.edu*; Zhi Qiao, University of Minnesota Rochester; Bo Hu, University of Minnesota Twin Cities; Maria Dominguez-Bello, University of Puerto Rico-Rio Piedras

Key Words: Helicobacter pylori, logistic regression, linear regression, gastric cancer, childhood asthma

The bacterium, Helicobacter pylori (H. pylori), colonizes the human stomach and is one of the factors contributing to gastric cancer, but prohibiting childhood asthma. In recent decades, the prevalence of H. pylori infection has been declining in developed countries. However, the approximate time when H. pylori will disappear and the quantitative association of the H. pylori declining rate with the trends of gastric cancer and childhood asthma have yet to be discovered. We take the data from the National Health and Nutrition Examination Survey III and the Surveillance Epidemiology and End Results databases; analyze the trends of H. pylori infection, gastric cancer and childhood asthma in the United States population by applying a linear and a logistic regression model; and set up a mathematical model to show the relationship of the H. pylori extinction rate with the prevalence of gastric cancer and childhood asthma. We also perform a model validation to test whether or not these models are the best fit to the data. According to the regression models, H. pylori will approximately disappear in U.S. at 2037.

Body Mass Index And Waist-Hip Ratio Interaction With Cognitive Performance

✦Yinghua (Grace) Zhang, GlaxoSmithKline, 130 Descanso Dr., Unit 440, San Jose, CA 95134, *yinghua.y.zhang@gsk.com*

Key Words: Dementia, BMI, Cross-sectional

It is a cross-sectional study to determine whether body mass index (BMI) is independently associated with cognitive function in postmenopausal women and the relationship between body fat distribution as estimated by waist-hip ratio (WHR). The participants are eight thousand seven hundred forty-five postmenopausal women aged 65 to 79 free of clinical evidence of dementia who completed the baseline evaluation in the WHI hormone trial. Participants completed a Modified Mini-Mental State Examination (3MSE), health and lifestyle questionnaires, and standardized measurements of height, weight, body circumference, and blood pressure. Statistical analysis was performed of associations between 3MSE score, BMI, and WHR after controlling for known confounders. We find BMI was inversely related to 3MSE score; for every 1-unit increase in BMI, 3MSE score decreased 0.988 points (P< .001) after adjusting for age, education, and vascular disease risk factors. BMI had the most pronounced association with poorer cognitive functioning scores in women with smaller waist measurements. In women with the highest WHR, cognitive scores increased with BMI.

Is Transmission Of Leishmaniasis Outbreak In Afghanistan Geographically Varied?

✦Oyelola Abdulwasiu Adegboye, American University of Afghanistan, P.O. Box 458, Central Post Office, Durulaman Road, KABUL, International 5000 Afghanistan, *aadegboye@auaf.edu.af*

Key Words: Leishmaniasis, vector-borne, spatio-temporal, infection

Leishmaniasis is a common vector-borne disease and very important protozoan infection. It is contracted through bites from sand flies and can result in chronic and non healing sores. This mostly occurs on any exposed skin and can be disfiguring and painful. Leishmaniasis is the third most important vector-borne disease and Kabul is known to have the world's largest incidence of cutaneous leishmaniasis with an annually estimated 250,000 cases. The aim of our study is to estimate the incidence of leishmaniasis at the district/provincial level and explore what geographical constraints operate under the outbreak of leishmaniasis in Afghanistan. Spatial transmission models will be used to investigate the spatio-temporal spread of the outbreak and spatial and non-spatial variability of the risk of leishmaniasis outbreak will be investigated.

Comparing Splines And Fractional Polynomials In Modeling Time-To-Event Data

✦Leila D. Amorim, Federal University of Bahia, Rua Joaquim Ferraro do Nascimento, 102, apto.1003,, Pituba, Salvador-Bahia, International 41830-440 Brazil, *leiladen@ufba.br*; NÌvea B. Silva, Federal University of Minas Gerais

Key Words: splines, fractional polynomials, cox model, time-to-event data, survival analysis

Smoothing approaches have become increasily popular in applied research, particularly due to their flexibility to model the functional form of continuous covariate's effect on an outcome. Splines methods and

fractional polynomials have extensively been used in various applications. Research has been conducted for evaluating the performance of such approaches on the fit of generalized linear models and Cox models. However, comparison between the smoothing approaches in the context of analyzing time-to-event data is still limited. We compared regression splines and smoothing splines, which are techniques that differ in the number of knots used, in the way of determining their positions and on how the parameters for the polynomial pieces are estimated, to the fractional polynomials. These approaches are applied to the analysis of diarrhea time-to-event data. Simulation studies examined the strengths and weakness of different smoothing approaches.

Comparing Alternative Psychiatric Diagnostic Criteria Sets Using Item Response Theory With Total Information Curves

✦Melanie M Wall, Columbia University Departments of Biostatistics and Psychiatry, 1051 riverside drive, unit 48, New York, NY 10032, *mmwall@columbia.edu*; Deborah S Hasin, Columbia University Department of Psychiatry - Substance Dependence Research Group

Key Words: item response theory, latent variables, measurement, substance abuse

IRT models are useful for assessing the ability of criteria sets to measure underlying psychiatric disorders, e.g. substance use disorders. IRT assumes the underlying psychiatric disorder is inherently continuous ranging from low to high severity and estimates how well each observable criterion measures different points along that continuum. The total information curve summarizes the combined criteria and provides a visual inspection of where along the disorder continuum the criteria measure the disorder well (i.e. with more precision). In the current work, we develop a way to statistically test whether the total information provided by one criteria set is larger than another. Standard errors for the total information are developed based on asymptotic theory as well as using the bootstrap procedure. We illustrate the method with an application to nicotine use disorder. With the changes proposed for the other DSM-5 substance use disorders to combine dependence and abuse criteria along with cravings into a single disorder, we examined whether combining nicotine dependence, abuse and cravings criteria is warranted.

Using Bayesian Logistic Regression To Estimate The Risk Or Prevalence Ratio

Charles Rose, Centers for Disease Control; ✦Andrew L. Baughman, Centers for Disease Control and Prevention, 1600 Clifton Road NE, MS E-30, Atlanta, GA 30329 USA, *alb1@cdc.gov*

Key Words: Risk Ratio, Prevalence Ratio, Bayesian Logistic Regression, Log-Binomial, Poisson Regression

In cohort and cross-sectional studies or when the outcome is common, the risk ratio (RR) is the preferred measure of effect rather than an odds ratio (OR). The logistic regression OR is often used to approximate the RR when the outcome is rare. However, whether the outcome is rare or common, logistic regression predicted exposed and non-exposed risks can be used to form an appropriate RR. We developed a Bayesian logistic regression model to estimate the RR, with associated credible interval, and applied the model to published data. We compared our results to four commonly used RR modeling techniques: stratified Mantel-Haenszel, logistic regression, log-Binomial, and log-Poisson. Our Bayesian logistic regression provides a flexible framework for investigating confounding and effect modification on the risk scale and compares favorably with existing RR modeling methods.

Detecting Population Substructure In Rare Variants And Common Variants Data

✦Dandi Qiao, Harvard University, Apt 134A, 199 Park Drive, BOSTON, MA 02215 US, *qddwudan@gmail.com*

Key Words: Population substructure, stratification, rare variants

In population-based association studies, population substructure gives rise to false positive results. Existing methods to adjust for population substructure are designed mostly to detect global stratification in common variants analysis. There is no formulated way to detect local stratification for either common variants or rare variants analysis. In this paper, we propose a novel test for population-based data to detect outliers and local population stratification in both common variants and rare variants analysis. We show by simulation that under few assumptions, this test has adequate type-I error and sufficient power in detecting global and local population stratification with both common variants data and sequence data. Also, we show that the test could be utilized in detecting subjects with unacceptable genotyping quality. We illustrate the test by applying it to a real datasets including German and Dutch subjects.

A Resampling-Based Approach For Controlling Type I Error In A Genomic Microarray Experiment In Which Case/Control Status And Chip Effect Are Confounded

✦Stephen Erickson, University of Arkansas for Medical Sciences, Arkansas Center for Birth Defects Research, 13 Children's Way #512-40, Little Rock, AR 72202 USA, *serickson@uams.edu*

Key Words: microarrays, genomics, methylation, resampling, permuation test

In microarray experiments, chip-to-chip measurement bias can be a potential source of false associations if case/control status and other key covariates are not properly balanced or randomized in the experimental design. This has not always been widely appreciated, nor always mentioned in the protocols of commercially produced microarrays. Thus, it is not uncommon for a statistician to be presented with a dataset in which chip effect is completely or partially confounded with case/control status. To properly control type I error in an experiment using the Illumina HumanMethylation27 microarray (Chowdhury et al 2011), we implemented a resampling-based approach in which an empirical null distribution of the t-statistic was generated by randomizing case/control status. The specific randomization procedure makes the null hypothesis artificially true, but retains the imbalanced experimental design and therefore the chance for spurious associations due to chip effect. Thus, the empirical null has heavier tails than the theoretical t distribution, which reduces power but properly controls type I error. In simulations, we estimate the reduction in power over a range of parameters.

Competing Risks Modeling In The Presence Of Left Censored Observations

✦Yushun Lin, University of Kentucky, KY 40536, *yushunlin@uky.edu*; Richard J. Kryscio, University of Kentucky

Key Words: competing risks, cumulative incidence function, Weibull regression model, proportional subdistribution hazards, left censored events

Death is a competing risk encountered when following a cohort of elderly subjects to dementia. In this manuscript we investigate the use of cumulative incidence functions (CIF) to analyze competing risks data. We propose an iterative Expectation-Maximization (EM) algorithm for nonparametrically estimating CIFs when one of the events (dementia) is subject to left and right censoring. To accommodate covariates in this situation the parametric proportional subdistribution hazard model due to Jeong and Fine (2007) is modified to assure that the sum of the underlying CIFs never exceeds one. Simulation studies support these modifications. An application to investigate the effect of genetics and education on the occurrence of dementia before death in the NUN Study is used to illustrate these results.

Using The Delta Method To Construct Confidence Intervals For The Odds Of Dementia Before Dying In A Nonhomogeneous Markov Transition Model

✦Liou Xu, University of Kentucky, , *loxu2@email.uky.edu*; Richard J. Kryscio, University of Kentucky; Lei Yu, Rush Alzheimer's Disease Center

Key Words: Multi-state Markov Chain, Nonhomogeneous, Dementia, Competing Event, Shared Random Effect, the Delta Method

In the study of chronic diseases like Alzheimer's, it is commonly the case that the investigators are particularly interested in the probability of disease onset before dying given a set of risk factors such as age, education, and genetic status. Our purpose is to continue the study for an extended nonhomogeneous Markov transition model that involves time dependent risk factors as well as the survival component, in which case the underlying transition probability matrix is no longer stationary. We focus on addressing the problem by accommodating the residual life time of the subject's confounding in the model. The convergence status of the chain is examined and the formulation of the absorption statistics is then derived. Based on the asymptotic normality of the sampling distribution, we propose using the Delta method to estimate the variance terms for the odds of dementia before dying for construction of confidence intervals. The results are illustrated with an application to the Nun Study data (Snowdon, 1997) in detail.

Attrition-Considered Statistical Characterization Of 15-Year Longitudinal Cognitive Change

✦Maria Josefsson, Department of Statistics, Umea University, Umea, SE-90187 Sweden, *maria.josefsson@stat.umu.se*

Key Words: cognitive decline, attrition, pattern mixture model

Episodic memory performance declines with the passage of time. Little is known about inter-individual differences in rate of change. Many longitudinal studies suffer from attrition as the mechanism causing the missing data often is non-ignorable. Another issue to consider when analyzing data from longitudinal studies including scores from tests, is the upper and lower boundaries of the test scores which limit the rate of change to move freely. Statistical methods not taking these issues into account can cause biased estimates. In order to handle the influence of missing data, we propose a random-effect pattern mixture model which also takes the ceiling/floor effects from the test scores into consideration. The model has been applied on an ongoing longitudinal study of episodic memory with 1444 participants identifying changes in episodic memory performance over time. Based on the rate of change and initial test score and factorized into different age groups, the episodic memory curve of the participants were categorized as successful, normal, or declining. In a second step, the "successful" individuals were characterized with regard to demographic, genetic, and lifestyle factors.

The Probabilistic Distribution Of Time To Relapse After Quitting Cigarette Smoking

✦Lei Li, RTI, 3040 Cornwallis Road, RTP, NC 27709, *lei@rti.org*

Key Words: smoking, survival analysis, mixture Weibull function

To estimate the probability of relapsing after having quitted cigarette smoking we analyzed the data from the 1998-1999 Tobacco Use Supplement to the Current Population Survey on quitting smoking among former smokers who reported having smoked at least 100 cigarettes in their lifetime and smoked daily. The Kaplan-Meier survival functions were first plotted for the time to relapse and then the mixture Weibull survival functions of the form S(t) = p x exp(-mtk) + (1-p), where (1-p) is the assumed proportion of permanent quitters, were used to fit the probabilistic distribution of the time to relapse by age and gender groups. The maximum likelihood estimates of the parameters m, k, and p were obtained via the EM algorithm. The estimated mixture survival functions were close to the Kaplan-Meier estimates. The hazard rate of relapsing was clearly not a constant due to the likely presence of a fraction of permanent quitters. Most of the relapses occurred within the first 100 days since quitting smoking and the relapse rate was nearly negligible beyond 200 days of abstinence.

Flexible Assessment Of Skewed Exposure In Case-Control Studies With Case Specific And Random Pooling

✦Neil J Perkins, NICHD / NIH, 6100 Executive blvd, rockville, MD 20852, *perkinsn@mail.nih.gov*; Brian Whitcomb, UMASS; Robert H Lyles, Rollins School of Public Health, Emory University; Enrique F. Schisterman, Eunice Kennedy Shriver, National Institute of Child Health and Human Development

Key Words: gamma distribution, set-based regression, pooling, biomarkers, efficient design

Pooled and hybrid, pooled-unpooled, designs have been proposed for epidemiologic study of biomarkers to minimize cost, while maintaining efficiency as well as several other physical and statistical benefits. Set-based logistic regression has been proposed for use with pooled data; however, analysis has been limited by assumptions regarding exposure distribution and logit-linearity of risk (i.e., constant odds ratio). We have developed a more flexible model for analysis of pooled or hybrid data using characteristics of the gamma distribution. A modified logistic regression is used to accommodate non-linearity corresponding

to removal of the restriction of equal shape parameters in the gamma distributed exposure for cases and controls. Full maximum likelihood estimation is compared to a more standard logistic regression approach via simulation to assess consistency and efficiency of risk effect estimates given random and disease specific pooled, hybrid data. Our methods are demonstrated through effect estimates of pooled cytokines data on perinatal outcomes.

Practical Considerations In Meta- And Pooled Analyses Of Type 2 Diabetes Data In Normal Weight Adults

✦Peter de Chavez, Northwestern University, 680 N. Lake Shore Drive, Suite 1400, Chicago, IL 60611, *p-chavez@northwestern.edu*; Juned Siddique, Northwestern University; Alan R. Dyer, Northwestern University; Mercedes R. Carnethon, Northwestern University

Key Words: meta-analysis, pooled analysis, diabetes, mortality, obesity, normal weight

Type 2 diabetes (T2DM) in normal weight adults is an intriguing representation of the metabolically obese normal weight phenotype. Since its prevalence is low (5-15% of all cases), no single study has had an adequate sample size to assess the health consequences of normal weight T2DM. However, several cardiovascular studies that include common measurements of variables relevant to T2DM have been conducted. Consequently, it is possible to carry out pooled and meta-analyses to test the hypothesis that mortality is higher among overweight persons with T2DM than persons with normal weight T2DM. Conducting pooled and meta-analyses presents a number of challenges such as non-overlapping demographic characteristics of cohorts (e.g., studies restricted to a given race or age); and the use of different measurement units, assays, and survey methodologies. This poster presents several methods used to combine results from 5 cardiovascular studies of 2,619 participants in order to estimate the effect of weight on mortality among adults with T2DM. We also describe the results of our analyses and discuss the limitations of our methods.

Development Of Neural Network Model To Predict Deoxynivalenol (Don) In Barley Using Forecasted Weather Conditions

✦Krishna Deepthi Bondalapati, South Dakota State University, Box 2108, Plant Scient Department, Brookings, SD 57007, *krishna.bondalapati@sdstate.edu*; Jeff Stein, South Dakota State University; Kathleen Baker, Western Michigan University

Key Words: Fusarium Head Blight, Scab, Neural Network, Mycotoxins, DON, Barley

Fusarium head blight of barley, caused by the fungus Gibberella zeae (anamorph: Fusarium graminearum), is a devastating disease in U.S. Northern Great Plains. Losses occur through the blighting of florets, disruption of grain fill, and most importantly through the contamination of grain with mycotoxins, primarily deoxynivalenol (DON). A weather-based predictive model has been developed for estimating the risk of economically significant DON accumulation in barley grain, however this model utilizes environmental conditions that have already occurred and therefore can only facilitate reactive disease management. Such a system is useful, but does not fully support integrated crop disease management strategies. In this research, a single layer neural network (NN) model was developed using a combination of measured and forecasted weather data to predict the risk of economic DON levels 5-days in advance to the period of peak crop susceptibility. The developed NN model had a prediction accuracy of 91% with a sensitivity of 76% and specificity of 96%. The model resulted in 89% prediction accuracy when tested on 55 independent field data collected from locations in the region of interest.

Genetic Variance Components Estimation For Binary Traits Using Multiple Related Individuals

✦Charalampos Papachristou, University of theSciences, 600 S 43rd St, Mailbox 64, Philadelphia, PA 19104, *c.papach@usp.edu*; Mark Abney, University of Chicago; Carole Ober, University of Chicago

Key Words: Binary Trait, Genetic Variance Components, GLMMs, MCEM, Diabetes, Complex Pedigrees

We propose a likelihood approach, developed in the context of generalized linear mixed models, for modeling dichotomous traits based on data from hundreds of individuals all of whom are potentially correlated through either a known pedigree, or an estimated covariance matrix. The advantage of our formulation is that it easily incorporates information from pertinent covariates as fixed effects and at the same time it takes into account the correlation between individuals that share genetic background or other random effects. The high dimensionality of the integration involved in the likelihood prohibits exact computations. Instead, an automated Monte Carlo expectation maximization algorithm is employed for obtaining the maximum likelihood estimates of the model parameters. Through a simulation study we demonstrate that our method can provide reliable estimates of the model parameters for sample sizes close to 500. Implementation of our method to data from a pedigree of 491 Hutterites evaluated for Type 2 diabetes (T2D) reveals evidence of a strong genetic component to T2D risk, particularly for younger and leaner cases.

Drug-Related Suicide Attempts: Increasing Disparities By Gender

✦Dhuly Chowdhury, RTI International, 6110 Executive Blvd, Suite 902, Rockville, MD 20852, *dchowdhury@rti.org*; Victoria Albright, RTI International ; Karol Krotki, RTI International

Key Words: DAWN, drug-related suicide attempts, SAMHSA

In the United States about every 15 minutes a person dies due to suicide, the fourth leading cause of death. This paper examines the number of drug-related suicide attempts (DRSAs) by gender and geographic area. Results from the Drug Abuse Warning Network (DAWN), an ongoing national public health surveillance system that monitors drug-related medical emergencies of non-Federal hospitals with 24-hour emergency departments, shows that in 2009 the estimated national male and female rates for DRSAs were 51.5 and 77.4 per 100,000. However, the gender difference varies by geographic area; this difference is significant for some areas; as a result the national level estimates are different. This paper will examine the DRSAs by gender for the nation and 13 metropolitan areas and will identify the areas where the female DRSAs are significantly higher compared to the national average for female. We will also identify the most frequently used drugs involved in DRSAs. Results may indicate the need to implement programs that

target specific geographic areas and potentially save lives. Furthermore, the difference between male and female rates for different age groups will be described.

A Simple Approach For Sample Size And Power Calculations For Clustered Count Data In Matched Cohort Studies

✦Dexiang Gao, University of Colorado Denver, 80045, *dexiang.gao@ucdenver.edu*; Gary Grunwald, University of Colorado Denver; Stanley Xu, Institute for Health Research, Kaiser Permanente Colorado

Key Words: Matched cohort design, Clustered count data, Random cluster effects Poisson model, Sample size, Statistical power

In matched cohort studies treated and untreated individuals are matched on certain characteristics to form clusters (strata) to reduce potential confounding effects. In this study design clustered count data are often the outcomes being used to estimate the treatment effect. Random cluster effects Poisson models (RCP) are frequently used for analyzing the clustered count data. However, sample size and power calculation can be challenging because the within cluster correlation needs to be considered but generally is not available in planning phase of a study. In this paper we compare the treatment effect estimate and its variance from RCP and those from other models. We then propose a simple approach for calculating statistical power and sample size for clustered count data in matched cohort studies with a constant matching ratio. Preliminary results from simulations showed that the power and sample size calculations are accurate when the random cluster effects are from either normal or gamma distributions. We also evaluated the simple approach of power and sample size calculations when the matching ratio is varying across strata.

From Electrocardiographic Data To Parameters Of Heart Rate Variability In The Study Of Workplace Stress Exposure And Cardiovascular Disease In Police Officers

Shengqiao Li, Centers for Disease Control and Prevention; ✦Anna Mnatsakanova, Centers for Disease Control and Prevention, 1095 Willowdale Rd, MS4050, Morgantown, WV 26505, *fma8@cdc.gov*; Cecil M Burchfiel, Centers for Disease Control and Prevention; James E Slaven, Indiana University School of Medicine; Luenda E Charles, Centers for Disease Control and Prevention ; John M John M. Violanti, University of Buffalo; Diane B Miller, Centers for Disease Control and Prevention ; Michael E. Andrew , Centers for Disease Control and Prevention

Key Words: Heart rate variability, Stress, Cardiovascular disease, Police health, HRV, ECG

The loss of heart rate variability (HRV) is believed to be a significant and specific marker of dysfunction in the system that allows individuals to return to normal levels of physiological arousal after exposure to a stressor. This dysfunction in the stress response system is brought about by a number of factors, particularly highly intense and chronic stress. Police work produces these conditions. Furthermore, lower HRV has been associated with increased levels of psychological distress, cardiovascular disease (CVD), and metabolic disorders and is thought to provide the link between stress, psychological disorders and the development of CVD. If HRV provides this link then it should be associated concurrently with stress exposure, psychological disorders and subclinical CVD. The goals of this research are to examine the relationships between HRV, exposure to police work stressors and stress related disease. We will present the complex sequence of data pre-processing and analysis required to obtain parameters for studying HRV using short-term electrocardiographic data collected on 462 police officers from the Buffalo Cardio-metabolic Occupational Police Stress (BCOPS) study.

The Association Of Male Circumcision And Bacterial Vaginosis In A Longitudinal Study

✦Jin Huang, University of Alabama at Birmingham, RPHB 327, 1530 3rd Ave S, Birmingham, AL 35294-0022, *jhuang09@uab.edu*; Jeff M Szychowski, University of Alabama at Birmingham; Suzanne P Cliver, University of Alabama at Birmingham; Mark A Klebanoff, Nationwide Children's Hospital; William W Andrews, University of Alabama at Birmingham

Key Words: generalized estimating equations, sexually transmitted infection, bacterial vaginosis, circumcision

Bacterial vaginosis (BV) is a common and recurrent condition in sexually active women. However, its etiology and its role as a sexually transmitted infection (STI) are still unclear. Our purpose was to evaluate the association between circumcision of male sexual partners, frequently investigated in STI research, and BV occurrence. In the Longitudinal Study of Vaginal Flora, 3620 nonpregnant women were followed quarterly for 1 year and BV was assessed based on Nugent Score, vaginal pH, and Amsel criteria. We used generalized estimating equations to estimate the effect of circumcision on occurrence of BV in intervals of risk. Based on Nugent Score, BV appeared in 24% of eligible intervals: 23% with at least 1 uncircumcised male sex partner reported and 24% with exclusively circumcised partners reported. We observed no statistically significant association between BV occurrence and circumcision status (OR=0.94, 95% CI: 0.75-1.18). Similar results were seen for the other BV diagnostic criteria. The association was nearly identical after controlling for other factors including partner race and condom use. We found no evidence of association between circumcision and occurrence of BV.

Semiparametric Single Index Model In 1-M Matched Case-Crossover Studies

✦Chongrui Yu, Virginia Polytechnic Institute and State University, 24060, *yucr@vt.edu*

Key Words: Matched case-crossover study, Single index model, Penalized spline, Fisher-von Mises, Polar coordinates

We propose a semiparametric single index model to analyze the matched case-crossover study. Penalized splines are used to model the semiparametric function of the "index". We develop both frequentist and Bayesian approaches to fit the model. Two Bayesian methods are developed using two different priors: 1) the prior distribution of index is taken as Fisher-von Mises, and 2) the index vector is represented via its polar coordinates and we use a uniform prior for the direction parameters. Simulation results indicate our Bayesian methods provide some improvement over the frequentist method. We demonstrate our approaches using an epidemiological example of a 1-4 bi-directional case-crossover study.

Analysis Of The Deaths For Neotlasy In The Rio Grande Do Sul, Brazil

✦Angela Isabel dos Santos Dullius, Universidade Federal de Santa Maria, 97110-000 Brazil, *angeladullius@gmail.com*; Tiane Camargo, FundaÁ„o Lusiada; Mariane Camargo Priesnitz, Universidade Luterana do Brasil

Key Words: Model, Box-Jenkins, Neotlasies

The malignant neotlasies constitute an increasing problem of public health in the world. In this paper, the Box-Jenkins methodology was applied to analyze historical data of the total deaths' number for neotlasy in the Rio Grande do Sul - Brazil, the period from January 1998 to December 2010. The Box-Jenkins approach is presented three steps - Identification, Estimation, Checking, are reviewed. Analyzing the autocorrelation coeficients and partial autocorrelation coeficients, one concludes that the best model that represents the data was the SARIMA(1,1,0)(1,0,0)6. The model that it presented residual average quadratic error of 0,0784. This work will go to contribute with excellent information for the public agencies, which need a system of information that supplies such knowledge.

Exploring Genome-Wide Gene-Gene Interaction Of Parkinson'S Disease

✦Taye H Hamza, Wadsworth Center, New York State Department of Health, 150 New Scotland Ave, Albany, NY 12208 USA, *thamza@wadsworth.org*

Key Words: Genetic, interaction, Parkinson's

The etiology of Parkinson's disease(PD) is not fully understood. To date many risk genes have been identified that explain only fraction of the heritability(Hamza and Payami, 2010). PD may be influenced by genes, environmental factors and their interactions. In spite of the computational burden and methodological challenge, exploring gene-gene interaction may help to uncover the missing heritability and explain the genetic contributions to PD(Cordel. 2009). We performed a genome-wide search of gene-gene interaction of 811K-SNPs genotyped for 2000 case-control pairs(Hamza etal. 2010). We screened all possible pairs (811Kx811k) by testing the differences in allelic association of SNPs between cases and controls, followed by standard case-control interaction(Purcell etal. Hum. Gen. 2007). Assessing all pairs of interaction is feasible in reasonable time by breaking down the whole-genome into sets of SNPs and performing parallel computing. We detected 5 independent pairs of SNPs which reached P< 10-10. Top hit pairs did not exhibit strong main effects. The results should be replicated in an independent study. SNPs with no main effect should not be ignored in the interaction analysis.

Application Of Pathway-Pdt To Identify The Underlying Genetics Of Autism Spectrum Disorder

✦YoSon Park, Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, , *ypark@med.miami.edu*; Eden R Martin, University of Miami Miller School of Medicine; Michael A Schmidt, University of Miami Miller School of Medicine; Margaret A. Pericak-Vance, Hussman Institute for Human Genomics, University of Miami Miller School of Medicine; Ren-Hua Chung, Hussman Institute for Human Genomics, University of Miami Miller School of Medicine

Key Words: Autism, Statistical Genetics, Pathway Analysis

Pathway analysis is useful for identifying variants with modest effects, clustered in multiple genes in a pathway and for studying the relationship of genes in the same pathway for disease susceptibility. Studying genes in a pathway in association with the disease susceptibility extends the use of GWAS data, where single-marker analyses haven't been very successful at identifying the causal variants in our autism study. We developed a novel family-based pathway analysis tool, Pathway-PDT (Abstract 302938, JSM 2011), which uses raw genotype data and accounts for pedigree information in the statistics. When genotype data in families are available, Pathway-PDT can have more power than methods using p-values only. We applied Pathway-PDT to two independent autism GWAS family datasets provided by Hussman Institute for Human Genomics (HIHG) and Autism Genetic Resource Exchange (AGRE), testing pre-defined pathways from the Reactome, KEGG and Gene Ontology (GO) databases. The most significant pathway, the bicarbonate transport pathway (GO: 15701) showed nominally significant associations in the independent analyses (HIHG, P=0.049; AGRE, P=0.009) and in the combined analysis (P< 0.0002).

Estimating The Improvement In Performance By Combining Biomarkers In A Two-Phase Study Design

✦Aasthaa Bansal, University of Washington, Department of Biostatistics, F-600, Health Sciences Building, Box 357232, Seattle, WA 98195 USA, *abansal@uw.edu*; Margaret Sullivan Pepe, Fred Hutchinson Cancer Research Center

Key Words: two-phase design, biomarker combination, diagnosis, ROC, classification

BACKGROUND: When an existing marker, X, does not have sufficient diagnostic accuracy on its own, a new marker Y may in combination with X improve performance. We investigate estimation of the increment in classification performance due to Y from a two-phase study with X measured for all subjects in a cohort and Y measured for a case-control subset. Additionally, we consider matching controls to cases within the case-control subset and explore its effect on the estimation of incremental value. METHODS: We develop non-parametric and semi-parametric methods for estimation of the ROC curve, as well as some more recently-developed measures such as the net reclassification index, integrated discrimination improvement (Pencina et al., 2008) and total gain (Bura & Gastwirth, 2001). We evaluate our methodology under both unmatched and matched case-control study designs using simulation studies. RESULTS: The methods are easy to implement and provide valid inference. Our results suggest that in most settings, efficiency of the estimation of incremental value is improved by matching controls to cases and by using simple semi-parametric methods for estimation.

190 Section on Statistics in Sports Speaker with Lunch (fee event)

Section on Statistics in Sports

Monday, August 1, 12:30 p.m.–1:50 p.m.

Measuring Fielding In Baseball: Present And Future

✦Shane Jensen, The Wharton School, , *stjensen@wharton.upenn.edu*

Fielding ability remains a difficult quantity to estimate in baseball. I present a sophisticated hierarchical model that uses current ball-in-play data to evaluate individual fielders. I will discuss continuing efforts to extend these fielding models to examine the evolution of fielding ability over multiple seasons. Many challenges in this area remain: our modeling efforts are constrained by the aspects of fielding measured in the current data. These limitations will be discussed with a look towards the potential availability of much higher resolution data in the near future.

191 Biopharmaceutical Section P.M. Roundtable Discussions (fee event)

Biopharmaceutical Section

Monday, August 1, 12:30 p.m.–1:50 p.m.

Implementation and Logistics of Adaptive Trial Design

✦Eva R Miller, ICON Clinical Research, 212 Church Rd, North Wales, PA 19454 USA, *Eva.Miller@iconplc.com*

Key Words: Adaptive Trial Design

Participants at this roundtable will discuss real world experiences in the implementation of adaptive trials. Statisticians who have actually conducted adaptive trials are welcome. as are statisticians who are about to conduct adaptive trials for the first time.. Topics include: roles and responsibilities of the unblinded statistician, communication paths to preserve trial integrity, data lag, working with several data sources, and managing randomization and drug supply within IVRS. Study teams experience greater demands for planning, communication and teamwork within adaptive trial designs than for traditional trials. We will discuss how these demands impact the statisticians, SAS programmers, Project Managers, data managers, and drug supply managers. Hopefully, participants will be better prepared to surmount the hurdles.

Statistical Issues in the Design and Analysis of Dose-Response Studies

✦Susan Huyck, Merck Research Laboratories, 126 East Lincoln Ave, Mail Stop: RY34-A240 PO Box 2000, Rahway, NJ 07065, *susan.b.huyck@merck.com*

Key Words: Dose Response, Design, Analysis

Dose response studies present a unique set of opportunities and challenges in terms of design and analysis. This roundtable discussion will focus on the statistical issues and approaches that may be found in dose response studies such as sample size estimation, selection of dose levels, trend tests vs. pairwise comparisons, use of adaptive designs and interim analyses, selection of optimal dose for phase III, etc.

Effect of Subsequent Therapies After Discontinuation of Study Medication in Analyzing Overall Survival in Cancer Clinical Trials

✦Julie Xiuyu Cong, Boehringer Ingelheim Pharmaceuticals Inc, 900 Ridgebury Rd, Ridgefield, CT 06877 USA, *julie.cong@boehringer-ingelheim.com*

Key Words: Subsequent treatment/therapy, treatment discontinuation, overall survival, clinical trials, cancer/oncology

Typically in cancer trials, patients are discontinued from the randomized study medication (control or experimental treatment) if patients develop disease progression or intolerable adverse events. In many cancer settings, such patients continue living for a considerably long time and take other anti-cancer therapies available to them. These patients are typically followed up for vital status in the trial with the aim to evaluate overall survival. As a result of such subsequent therapies, the treatment effect of true interest (experimental vs. control) is not properly assessed using the standard Intent-to-Treat (ITT) approach. Various statistical methodologies have been proposed and investigated by statisticians worldwide, including but not limited to marginal structural model, nested structural model, propensity scores, etc. This roundtable session aims to discuss the pros and cons of available methodologies and stimulate new ideas.

Analysis Of Longitudinal Categorical Data

✦Madhuja Mallick, Merck Research Laboratories, , *madhujamallick@gmail.com*

Objective of longitudinal data model is to study changes over time within subject or changes over time between subjects. When the longitudinal data is categorical, modeling the data is not easy. Since multivariate normality assumption does not hold. The issue becomes more complicated when there are missing data. Strategies to deal with missing data affect the inference procedure as well. This session will focus on different candidate models such as marginal, conditional, and transitional models and their advantages and disadvantages in presence of missing data. Handling this type of data in optimal way in terms of operation characteristics will also be of interest for this session.

CANCELLED: Impact Of The Non Inferiority Clinical Trials Draft Fda Guidance On Non Inferiority Margins Interpretation

Key Words: non inferiority, guidance, sample size

In March 2010, the FDA has issued this guidance for comments.This guidance introduces the M1 and M2 non inferiority margin definitions and compares the 95-95 fixed margin method to the synthesis method. M1 is defined as the entire effect of the active control assumed to be present in the NI study and M2 as the largest clinically acceptable difference of the test drug compared to the active control. Ruling out a difference between the active control and test drug larger than M1 is the critical finding that supports a conclusion of effectiveness (small p-value requestd). However, as M2 represents a clinical judgment, there may be a greater flexibility in interpreting a 95% upper bound for C-T that is slightly greater than M2, as long as the upper bound is still well

less than M1. The objective of the roundtable is to investigate whether this guidance has changed the current statistical practise related to sample size calculation and the way to interpret NI clinical trials.

Health Care Reform, Baby Boomers, And The Pharmaceutical Industry

✦T. Paulette Ceesay, Merck and Comapny, 351 N. Sumneytown Pike, North Wales, PA 19454, *paulette_ceesay@merck.com*; Darcy Hille, Merck and Company

Key Words: healthcare reform, baby boomers, pharmaceutical industry

The Patient Protection and Affordable Care Act or healthcare reform is the farthest-reaching piece of social welfare legislation in four decades and is still a hotly debated topic. Some members of the general public are still confused over its provisions. The first Post World War II baby boomers are beginning to turn 65 and many of them have been hit hard by the recession with job cuts and catastrophic losses to their assets and 401ks. Government may seek to limit the amount of funds spent on prescription drugs to increase availability to a larger population resulting in placing a heavier burden on consumers for the purchase of cutting edge drugs, therapies and patient specific prescriptions due to non-responsiveness to generics. The discussion will focus on how we as statisticians may be able to predict the impact of this historic legislation to the pharmaceutical companies' bottom line as it relates to cost, research and development.

192 ENAR P.M. Roundtable Discussion (fee event)

ENAR

Monday, August 1, 12:30 p.m.–1:50 p.m.

Methodology To Assess Surrogate Markers In Clinical Trials

✦Jeremy Michael George Taylor, University of Michigan, Dept of Biostatistics, 1420 Washington Heights, Ann Arbor, MI 48109, *jmgt@umich.edu*

There is a strong desire to use surrogate endpoints in randomized clinical trials, instead of true clinical endpoints. Using surrogate endpoints could shorten the duration of the trials and reduce their cost, enabling more therapies to be evaluated. There are significant statistical challenges in demonstrating that a potential surrogate endpoint is valid. A number of statistical approaches have been suggested, including metrics for the proportion of treatment effect explained, within and between trial measures of association between the surrogate and the true endpoint, and counterfactual models.

193 Section for Statistical Programmers and Analysts P.M. Roundtable Discussion (fee event)

Section for Statistical Programmers and Analysts

Monday, August 1, 12:30 p.m.–1:50 p.m.

The Future of U.S. Statistical Programmers in Pharmaceutical Industry

✦Chengying (Nancy) Wu, Sanofi-aventis, , *nancy-gv.wu@sanofi-aventis.com*

Key Words: Statistical Programming, Outsourcing, SAS programming, Pharmaceutical Industry

With the big waves of moving in Pharmaceutical Industry Research and Development due to this deep recession, more and more statistical programming jobs are shifted to oversea from United States by operating more R&D sites or outsourcing. What is our future as a Statistical programmers in the United States? What do we need to prepare to win this challenge? Bring your questions, thoughts, or suggestion in and benefit from others'!

194 Section on Bayesian Statistical Science P.M. Roundtable Discussion (fee event)

Section on Bayesian Statistical Science

Monday, August 1, 12:30 p.m.–1:50 p.m.

Manifold Learning For High-Dimensional Data

✦David Dunson, Duke University, Durham, NC 27708, *dunson@stat.duke.edu*

Key Words: Bayesian, Machine learning, Dimensionality reduction, Nonparametric, Data fusion, Functional Data Analysis

There is increasing interesting in developing methods for compression, analysis and interpretation of massive dimensional data including not just vectors of continuous variables in a Euclidean space (e.g., gene expression) but also discrete data (e.g., gene sequences) and more complex objects such as functions, images, documents and movies. In addition, one often desires a joint representation and analysis of high-dimensional data of varying modalities. For example, for a patient in the emergency department, one may have data from diagnostic tests consisting of images and curves, while also having text from physician notes and categorical and continuous predictors, with the goal being to diagnose the condition and recommend an optimal treatment based on this disparate combination of data. To address such problems, methods of dimensionality reduction and joint modeling are needed. One direction to take is to suppose that the massive-dimensional observed data are concentrated near a (much) lower dimensional manifold. By "learning" this manifold, one can potentially enable compression of the data leading to dramatic storage, processing and analysis speed-ups. In this round-table

195 Section on Health Policy Statistics P.M. Roundtable Discussion (fee event)

Section on Health Policy Statistics
Monday, August 1, 12:30 p.m.–1:50 p.m.

Beyond Propensity Scores: What Could Be Finer?

✦Robert Obenchain, Principal Consultant, Risk Benefit Statistics LLC, 13212 Griffin Run, Carmel, IN 46033-8835, *wizbob@att.net*

Key Words: observational studies, propensity scores, patient subgroups, clustering patients, conditional independence, local control

In their classic 1983 paper in Biometrika, Rosenbaum and Rubin (R&R) established that conditioning on a true Propensity Score (PS) causes the joint distribution of patient x-covariates and t-treatment assignments to factor into two independent terms. One factor shows that patients have thereby been conditionally blocked on their x-distributions, while the other term quantifies conditional balance / imbalance on the fractions of patients receiving the two treatment choices. Furthermore, R&R established that the true PS is the "most coarse" of all possible such factoring scores, while the vector of patient-level x-covariates provides the "most fine" factoring score. Our discussion will address the question: "Are there factoring scores between the true PS and the individual x-vectors, and what would be their practical advantages in adjustment for treatment selection bias and confounding in observational studies?"

Data Confidentiality: Health Policy Perspective

✦Ofer Harel, University of Connecticut, 215 Glenbrook Road U-4120, Storrs, CT 06269, *oharel@stat.uconn.edu*

Key Words: confidentiality, privacy, synthetic data

There is increasing demand for access to economic, medical, educational, and human services data while at the same time growing concerns about confidentiality for the individuals whose information is being collected. Often times it is unethical to release private information to the public, and for certain medical and educational data, it is illegal. Currently, in many situations, researchers have to go through a lengthy process to gain access to such data and, even after being approved, are only granted access to limited data sets or data that have been perturbed in some way. Although this may protect individuals' privacy and confidentiality, it may severely limit the utility of the information. Somehow, a balance must be struck between the release of data for research purposes and the risk of disclosing private information. In this roundtable discussion we will discuss the conceptual, methodological and public policy implications of these issues and how these issues augment exciting challenges and provide open problems for our field.

196 Section on Quality and Productivity P.M. Roundtable Discussion (fee event)

Section on Quality and Productivity
Monday, August 1, 12:30 p.m.–1:50 p.m.

Quality Excellence In Design And Manufacturing: A Roadmap To Customer Delight

✦Daksha Chokshi, Pratt & Whitney Rocketdyne, M/S 731-84, P.O. Box 109600, West Palm Beach, FL 33410, *daksha.chokshi@pwr.utc.com*

Key Words: Design for Six Sigma (DFSS), Axiomatic Design, Quality Function Deployment (QFD), Simulations, Theory of Inventive Problem Solving (TRIZ)

Quality excellence in both Design and Manufacturing are keys to the success for any business in delighting the customers with products that meet or exceed their expectations. This roundtable will explore important linkages, protocols, and lessons learned from successful Manufacturing and Six Sigma Design applications. In particular, the cost of a design change made in the engineering phase prior to release to manufacturing is much lower than the same change if it is made after the design is released to manufacturing. The importance of using the right tools for the right applications in manufacturing and design will be stressed. We will also discuss understanding the influence that design choices have on achieving a robust manufacturing system.

197 Section on Statistical Computing P.M. Roundtable Discussion (fee event)

Section on Statistical Computing
Monday, August 1, 12:30 p.m.–1:50 p.m.

To Gui Or Not To Gui: R In Intro Biostats For Biomedical Graduate Students

✦Hao Liu, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030 USA, *haol@bcm.edu*; Susan Galloway Hilsenbeck, Baylor College of Medicine

Key Words: R, GUI, introduction to statistics, class, teaching

We teach an introductory biostatistics course in the Translational Biology and Molecular Medicine graduate program. In response to student requests for software tools that they can use in their further graduate and medical school work, we began using R about 4 years ago as the software/homework platform. We considered other commercial options and settled on R, in part because of its wide capabilities, availability for PC's and Mac's, and favorable pricing (i.e. free). In the class, we first introduce command-line R, but then use both RCommander, as a GUI interface, and scripts, for worked problems and homework assignments. In this roundtable, we will share pros and cons of using R in this venue, hopefully hear other experience with R GUI's, exchange perspectives on introducing R to PhD and MD students with little programming background, and share suggestions from student feedback.

198 Section on Statistical Consulting P.M. Roundtable Discussion (fee event)

Section on Statistical Consulting
Monday, August 1, 12:30 p.m.–1:50 p.m.

Documenting, Tracking, And Recognizing Consulting Effort In An Academic Medical Center

✦Matthew Stuart Mayo, University of Kansas Medical Center, MS 1026, 3901 Rainbow Boulevard, Kansas City, KS 66160, *mmayo@kumc.edu*

Key Words: In-kind support, Project management, E-Project

The need for documenting and tracking consulting effort of statisticians/biostatisticians in academic medical centers is ever increasing. Given these increased demands, recognition of these efforts needs to be incorporated by departments and institutions. This will be an open discussion, facilitated by examining web-based tools developed by the Department of Biostatistics at the University of Kansas Medical Center for documenting and tracking and how that information has improved the recognition of these efforts by the Institution.

199 Section on Statistical Education P.M. Roundtable Discussion (fee event)

Section on Statistical Education
Monday, August 1, 12:30 p.m.–1:50 p.m.

Tricks of the Trade

✦Megan Elizabeth Mocko, University of Florida, 2525 NW 50th Place, Gainesville, FL 32605 USA, *mmeece@stat.ufl.edu*

Key Words: statisical education, activities, engagement

Do you have an activity or example that you use in class that illustrates a point really well or engages the students? Do you want to have more? Are you willing to share? Then, this is the roundtable for you. Starting with the roundtable leader, each participant will be asked to share an activity or two that works well for them. By the end of meal, you should walk away with "a new bag of tricks" to try in the upcoming year.

Teaching English Language Learners in an Online Environment

✦Amy Elizabeth Wagler, The University of Texas at El Paso, 125 Bell Hall, 500 W. University Ave., El Paso, TX 79968, *awagler2@utep.edu*

Key Words: online classes, english language learners, statistics education

It is estimated that 1 in 9 U.S. Kindergarten through twelfth grade students are considered English Language Learners (ELLs). This proportion is projected to increase to 1 in 4 by 2030. Many universities are experiencing an analogous, though less dramatic, increase in the proportion of ELLs enrolled at their institution, and at the same time, are encouraging more and more courses to be offered online. In response to these trends, this roundtable will explore the issues faced by ELLs when learning statistical concepts in online environments. Ideas for assisting ELLs in online statistics courses will be proposed and critiqued. Reasons behind advantages for ELLs in online environments as opposed to F2F environments will also be discussed. Informed by qualitative and quantitative evidence about specific challenges faced by ELLs, some time will be spent brainstorming about pedagogical methods effective in addressing these challenges in an online environment.

200 Section on Statistical Graphics P.M. Roundtable Discussion (fee event)

Section on Statistical Graphics
Monday, August 1, 12:30 p.m.–1:50 p.m.

Visualizing High-Dimensional Data

✦Genevera I. Allen, Baylor College of Medicine & Rice University, Department of Statistics, 6100 Main St., MS 138, Houston, TX 77005 USA, *gallen@rice.edu*

Key Words: data visualization, high-dimensional data

High-dimensional data sets are common in areas of medicine, image analysis, climate studies, Internet tracking, and finance. Visualizing massive amounts of data can pose a major challenge. In this roundtable, we will discuss different approaches and recent developments for visualizing high-dimensional data sets. Particular points of discussion will include methods for finding and visualizing patterns, visually assessing the fit of a model, and letting graphics drive the data analysis.

201 Section on Statistics and the Environment P.M. Roundtable Discussion (fee event)

Section on Statistics and the Environment
Monday, August 1, 12:30 p.m.–1:50 p.m.

The Present and Potential Role of Statistics in Paleoclimate Reconstruction for Learning About Climate Change

✦Bo Li, Purdue University, 150 North University St., Department of Statistics, Purdue University, West Lafayette, IN 47907, *boli@purdue.edu*

Key Words: Climate change, Paleoclimate reconstruction, Proxies

Paleoclimate reconstruction plays an important role in understanding climate change. Typically, a reconstruction is based on climate information from proxies such as tree rings, pollen, boreholes or other observations. It poses a challenging statistical problem because the climate signal is often embedded in substantial noise. The scientific significance and statistical flavor has created a growing interest from statisticians. We will discuss the present and potential role of statistics in paleoclimate reconstruction for learning about climate change.

202 Section on Statistics in Epidemiology P.M. Roundtable Discussion (fee event)

Section on Statistics in Epidemiology
Monday, August 1, 12:30 p.m.–1:50 p.m.

Instrumental Variable Estimation In Epidemiologic Research

✦Miguel Hernan, Harvard School of Public Health, 677 Huntingtin Avenue, Boston, MA 02115, *mhernan@hsph.harvard.edu*

Key Words: instrumental variables, causal inference, epidemiology, confounding

Instrumental variable (IV) estimation is an attractive approach for causal inference from observational data. Unlike other approaches, IV estimation does not require that all confounders are appropriately measured and adjusted for in the analysis. Rather, valid IV estimation requires (i) the identification of an instrument, and (ii) that some additional conditions hold. Unfortunately, it is not generally possible to empirically verify whether a variable is an instrument or whether the additional conditions hold. Further the direction of bias of IV estimates may be counterintuitive for epidemiologists accustomed to other adjustment methods. In this Roundtable, we will discuss the relative advantages and disadvantages of IV estimation for causal inference from observational data.

Hot Topics In Genetic Epidemiology

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

Key Words: genetic epidemiology, statistical genomics, next generation sequencing, gene environment interaction, epigenetics, risk prediction

Great advances in genomic technologies in the past several years have transformed the field of genetic epidemiology with much more data to explore and many more possible research paths to explore. It has become a real possibility to study the effects of genes, environment, and their interplay on disease risk at the biological systems level based on data collected from complex study designs. The large, diverse, and affordable genomics and proteomics data pose many study design, statistical, and computational challenges, and have stimulated vigorous methodology and theory developments lately. In this roundtable, we will discuss what have been accomplished through technological revolutions, the many gaps remaining between biological questions to be addressed and available methodologies, and many collaborative and grant opportunities in genetic epidemiology.

203 Section on Survey Research Methods P.M. Roundtable Discussion (fee event)

Section on Survey Research Methods
Monday, August 1, 12:30 p.m.–1:50 p.m.

Measurement Error in Survey Paradata

✦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: Survey administration, Survey estimation, Data quality, Survey design, Survey operations

The use of paradata, or data collection process information, to enhance survey adminstration, survey estimation, and the monitoring of survey data quality has become almost ubiquitous with recent advances in data collection technology. Unfortunately, little is known about the error properties of the paradata themselves, and the implications of these error properties for survey operations and the quality of survey estimates. The objectives of this roundtable discussion will be to 1) discuss current uses of survey paradata, 2) brainstorm methods for quantifying the error in paradata collected using a variety of modes, 3) discuss study designs for analyzing the implications of the measurement error for survey operations and estimation, and 4) identify meaningful research agendas in this area.

204 Section on Teaching of Statistics in the Health Sciences P.M. Roundtable Discussion (fee event)

Section on Teaching of Statistics in the Health Sciences
Monday, August 1, 12:30 p.m.–1:50 p.m.

Using The Writings Of Mark Twain To Reduce Introductory Statistics Course Anxiety

✦Jack Barnette, Colorado School of Public Health, 13001 East 17th Place, MS B119, Room C3000D, Aurora, CO 80045, *jack.barnette@ucdenver.edu*

Key Words: course anxiety, statistics training methods, Mark Twain

For many the introductory statistics course is an anxiety producing activity. Aspects of the instructional environment and approach taken by the instructor influence this anxiety. This roundtable demonstrates how literature may be used to introduce and provide input for discussion of theory and methods. For over 30 years, the presenter has used the writings of Mark Twain to introduce statistical concepts. Few recognize the keen sense Twain had of the use of probability, prediction, and basic experimental design found in many of his writings. Twain presents examples, including: the use of probability theory in "Science vs. Luck", central tendency and variability in "My Watch", statistical inference in "The Danger of Lying in Bed", distribution theory in "Passage from a Lecture", and correlation and regression in "Life on the Mississippi." These provide anxiety relief for stressed students and appreciation from others and they provide discussion points about appropriate and inap-

propriate applications of theory and methods. Attendees will observe how these writings may be used in statistical training in a variety of education, social science, and health fields.

205 Social Statistics Section P.M. Roundtable Discussion (fee event)

Social Statistics Section

Monday, August 1, 12:30 p.m.–1:50 p.m.

Can Statisticians Add Value to Teacher Value-Added?

✦Daniel F McCaffrey, RAND, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA , *danielm@rand.org*

Key Words: mixed models, longitudinal data

Measuring teachers' performance using "value-added" estimated from their students' longitudinal achievement test score data is the center of education policy and controversy. It is the centerpiece of the $4 billion US Department of Education Race to the Top Grant Program and a $500 million study by the Bill and Melinda Gates Foundation, a part of growing number of state and school district accountability programs, and the source of controversy in Los Angeles and New York where local newspapers reported or went to court for the right to report individual teacher value added. What role are statisticians playing in the estimation of value-added? What are the key statistical issues in the estimation of value-added? And what role can our profession play in the coming years to make the data most useful for improving the education of our country's children?

206 Late Breaking Session: Heritage Health Prize

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

Monday, August 1, 10:30 a.m.–12:20 p.m.

Heritage Health Prize

✦Hadley Wickham, Rice University, 77004 USA, *hadley@rice.edu*; ✦Chris Volinsky, AT&T Labs-Research, , *volinsky@research.att.com*; ✦Genevera I. Allen, Baylor College of Medicine & Rice University, Department of Statistics, 6100 Main St., MS 138, Houston, TX 77005 USA, *gallen@rice.edu*; ✦Mark Culp, WVU, , *mculp@stat.wvu.edu*; ✦Noah Pepper, Lucky Sort, 8083 SE 13th Avenue #2, Portland, 97202, *noah@luckysort.com*; ✦Homer Strong, Lucky Sort, 8-83 SE 13th Avenue, #2, Portland, 97202, *homer@luckysort.com*; ✦Eric Chi, Rice University,

The session will begin with a brief introduction to the Heritage Health Prize by myself and Genevera and will continue with two talks giving insight into what the competition will reveal. It will conclude with two modern statistical approaches likely to be useful for modeling this type of data. * Introduction to the Heritage Health Prize Genevera Allen and Hadley Wickham * Lessons learned from the netflix prize Chris Volinsky * Privacy vs Efficacy: The Double-Edged Sword of Health Data Homer Strong and Noah Pepper * Robust parametric classification and variable selection by a minimum distance criterion. Eric Chi * Multi-view Learning with Boosting Mark Culp

207 Teaching an old dog new tricks: parallel, adaptive, and automated Monte Carlo methods appearing in JCGS ■●

JCGS-Journal of Computational and Graphical Statistics, International Indian Statistical Association, Section on Statistical Computing, Section on Statistical Graphics

Monday, August 1, 2:00 p.m.–3:50 p.m.

Divide and Conquer: A Mixture-Based Approach to Regional Adaptation for MCMC

✦Radu Craiu, University of Toronto, Department of Statistics, 100 St George Street, Rm 6010, Toronto, ON M5S 3G3 Canada, *craiu@utstat.toronto.edu*; Yan Bai, University of Toronto; Antonio Fabio Di Narzo, Swiss Institute of Bioninformatics

Key Words: Adaptive MCMC, Mixture Model, Multimodal Distribution, Online EM, Regional Adaptation

The efficiency of Markov chain Monte Carlo (MCMC) algorithms can vary dramatically with the choice of simulation parameters. Adaptive MCMC (AMCMC) algorithms allow the automatic tuning of the parameters while the simulation is in progress. A multimodal target distribution may call for regional adaptation of Metropolis-Hastings samplers so that the proposal distribution varies across regions in the sample space. In the case in which the target distribution is approximated by a mixture of Gaussians, we propose an adaptation process for the partition. It involves fitting the mixture using the available samples via an online EM algorithm and, based on the current mixture parameters, constructing the regional adaptive algorithm with online recursion (RAPTOR). The method is compared with other regional AMCMC samplers and is tested on simulated as well as real data examples.

The Polya Tree Sampler: Toward Efficient and Automatic Independent Metropolis-Hastings Proposals

✦Timothy Hanson, University of South Carolina, Department of Statistics, 216 LeConte College, Columbia, SC 29208, *hansont@stat.sc.edu*; Joao Monteiro, University of Minnesota; Alejandro Jara, Pontificia Universidad CatÛlica de Chile

Key Words: Markov chain Monte Carlo, Indepdence proposal, Bayesian nonparametrics, Polya tree

We present a simple, efficient, and computationally cheap sampling method for exploring an unnormalized multivariate density, such as a posterior density, called the Polya tree sampler. The algorithm constructs an independent proposal based on an approximation of the target density. The approximation is built from a set of (initial) support points -- data that act as parameters for the approximation -- and the predictive density of a finite multivariate Polya tree. In an initial warming-up phase, the support points are iteratively relocated to regions of higher support under the target distribution to minimize the distance

between the target distribution and the Polya tree predictive distribution. In the sampling phase, samples from the final approximating mixture of finite Polya trees are used as candidates which are accepted with a standard Metropolis-Hastings acceptance probability. Several illustrations are presented, including comparisons of the proposed approach to Metropolis within-Gibbs and delayed rejection adaptive Metropolis algorithm.

On the Utility of Graphics Cards to Perform Massively Parallel Simulation of Advanced Monte Carlo Methods

✦Anthony Lee, University of Oxford, Department of Statistics, Oxford, OX1 3TG UK, *lee@stats.ox.ac.uk*; Chris Holmes, University of Oxford; Arnaud Doucet, University of British Columbia; Chris Yau, University of Oxford; Mike Giles, University of Oxford

Key Words: GPU, stochastic simulation, Monte Carlo, MCMC, Particle Filter

For certain types of scientific algorithms, desktop graphics cards using graphical processing units (GPUs) offer the performance of cluster-based computing at a fraction of the cost. Moreover, GPUs are dedicated, low maintenance, energy-efficient devices that are becoming increasingly easy to program. In this talk we overview the class of statistical algorithms amenable to GPU computation. We then present a case study using advanced Monte Carlo algorithms including population-based MCMC, sequential Monte Carlo samplers and the particle filter. We demonstrate that GPUs can lead to substantial speedups ranging from 35 to 500 fold over conventional CPU single-threaded computation. This suggests that GPUs and other multi-core devices are likely to change the landscape of high performance statistical computing in the near future.

208 Controversies in Sports ■●

Section on Statistics in Sports

Monday, August 1, 2:00 p.m.–3:50 p.m.

The Tarzan Effect: Why Are Athletes Today Better Than in Years Past?

✦Scott Berry, Berry Consultants, 77845 USA, *scott@berryconsultants.com*

Key Words: Bayesian Modeling, Sports, Population Modeling

In 1928 Johnny Weissmuller, who later played Tarzan in movies, was a famous Olympic champion swimmer. If the 1928 Olympic champion was able to swim 15 consecutive 100 meter swims (a relay of 15 Tarzans), of his world record 100 meter time, he would lose to a single modern athlete swimming a 1500 meter race. In objectively measured sports, such as track and field, speed skating, and swimming, world records, and Olympic times are getting better. There are likely many reasons why. Surely it is a combination of better training and better health, but a misunderstood effect is clearly that the population is bigger. We create Bayesian models of the effect of the population on Olympic winning times and find that the population of the world explains the increase in Olympic times very well and can be used to explain much of the improvement in Olympic times.

The Bowl Championship Series: Still Crazy After All These Years

✦Hal Stern, Department of Statistics, University of California, Irvine, 6215 Bren Hall, Irvine, CA 92697-3425 USA, *sternh@uci.edu*

Key Words: football, championship, ratings, prediction

The college football national championship for the highest level of competition is decided by a complicated arrangement called the Bowl Championship Series (also known as the BCS) that has been developed over the last 15 years. It remains a source of great controversy, heightened by the recent release of a ball called ``Death to the BCS''. This talk reviews the role of statistical methods in the BCS and also considers what an alternative approach to choosing a champion might be like.

Estimating Steroid Effects Using Aging Curves

This paper addresses the problem of comparing abilities of players from different eras in professional baseball. We borrow from the Berry, Reese, and Larkey approach to nonparametric estimation of aging functions for homerun hitters in Major League Baseball. We use additive models to estimate the innate ability of players, the effects of aging on performance, and the relative difficulty of each year within a sport. We measure each of these effects separated from the others. Hierarchical models are used to model the distribution of players. We specify separate distributions for each decade, thus allowing the ``talent pool'' within each sport to change. Nonparametric aging functions are used to estimate player specific aging functions. We use these functions to assess the impact of steroids on performance.

209 Recent Advances in Finite Mixture Models and Clustering

International Indian Statistical Association, International Chinese Statistical Association, SSC

Monday, August 1, 2:00 p.m.–3:50 p.m.

Clustering of High-Dimensional Data via Factor Models

✦Geoffrey John McLachlan, University of Queensland, Department of Mathematics, St. Lucia, Brisbane, 4072 Australia, *gjm@maths.uq.edu.au*

Key Words: High-dimensional data, Clustering, Normal mixture models, Factor models, Number of clusters

There has been a proliferation of applications in which the number of experimental units n is comparatively small but the underlying dimension p is extremely large as, for example, in microarray-based genomics and other high-throughput experimental approaches. Hence there has been increasing attention given not only in bioinformatics and machine learning, but also in mainstream statistics, to the analysis of complex data in this situation where n is small relative to p. In this talk, we focus on the clustering of high-dimensional data, using normal mixture models. Their use in this context is not straightforward, as the normal mixture model is a highly parameterized one with each

component-covariance matrix consisting of p(p+1)/2 distinct parameters in the unrestricted case. Hence some restrictions must be imposed and/or a variable selection method applied beforehand. We shall focus on the use of factor models that reduce the number of parameters in the specification of the component-covariance matrices. We also consider the problem of assessing the significance of groups in high-dimensional data using a resampling approach.

Fast Inference for Model-Based Clustering of Networks Using an Approximate Case-Control Likelihood

✦Adrian Raftery, University of Washington, Department of Statistics, Box 354322, Seattle, WA 98195-4322, *raftery@u.washington.edu*; Xiaoyue Niu, University of Washington; Peter Hoff, University of WAshington; Ka Yee Yeung, University of Washington

Key Words: clustering, genome science, graph, Markov chain Mone Carlo, protein-protein interaction, social network

The model-based clustering latent space network model represents relational data visually and takes account of several basic network properties. Due to the structure of its likelihood function, the computational cost is of order O(n^2), where n is the number of nodes. This makes it infeasible for large networks. We propose an approximation of the log likelihood function. We adapt the case-control idea from epidemiology and construct an approximate case-control likelihood which is an unbiased estimator of the full likelihood. Replacing the full likelihood by the case-control likelihood in the MCMC estimation of the latent space model reduces the computational time from O(n^2) to O(n), making it feasible for large networks. We evaluate its performance using simulated and real data. We fit the model to a large protein-protein interaction data using the case-control likelihood and use the model fitted link probabilities to identify false positive links.

Projection Pursuit via White Noise Matrices

✦Bruce George Lindsay, Penn State University, 326 Thomas Building, University Park, PA 16802, *bgl@psu.edu*

Key Words: projection pursuit, white noise, exploratory data analysis, dimension reduction

Projection pursuit dates back to 1974 (Friedman and Tukey). The goal is to find the most interesting projections of high dimensional data, where interesting is often defined to mean non-normal. The word "pursuit" arises from the numerical challenge of finding best projections. In this paper, the most interesting projection is the one that has the least conditional normality. This definition is turned into a statistical method by the construction of a white noise matrix based on the Fisher information matrix, and using the result that the normal distribution satisfies a minimal information property. An eigen-analysis of the white noise matrix yields both the most interesting and least interesting projections as the eigenvectors corresponding to the largest and smallest eigenvalues. The matrix is called a white noise matrix because if a projection is "white noise", in the sense of being marginally normal and independent of all orthogonal projections, it is an eigenvector with zero eigenvalue. A computationally fast, and statistically robust, estimator of the information matrix is developed, and used to find interesting directions in historically important and new data sets.

Comparing Different Points of View for Analyzing Finite Mixture Models

✦Gilles CELEUX, INRIA, UniversitÈ d'Orsay, B,t. 425, Orsay, F91405 France, *Gilles.Celeux@inria.fr*

Key Words: Conditional Completed Likelihood, Integrated Completed Likelihood, Variational Bayes

Abstract: Mixture models are an efficient tool to deal with heterogeneity or for model-based cluster analysis. These two points of view could lead to different methods for statistical inference (parameter estimation and model selection). After a survey highlighting their differences, the consequences of those two points of view on statistical analysis will be discussed. On the other hand, in a Bayesian perspective, the differences between a Bayesian inference through MCMC and variational approximation will be discussed.

210 Cutoff Sampling in Federal Establishment Surveys: An Inter-Agency Review ■

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

Monday, August 1, 2:00 p.m.–3:50 p.m.

Cut-off Sampling in Federal Establishment Surveys: An Inter-Agency Review

✦Daniel R. Yorgason, Bureau of Economic Analysis, 1441 L Street, NW, Washington, DC 20230, *dan.yorgason@bea.gov*; Janice Lent, Energy Information Administration; Benjamin Russell Bridgman, Bureau of Economic Analysis; Yan K. Liu, Statistics of Income/IRS; Alan H. Dorfman, U.S. Bureau of Labor Statistics; Yang Cheng, U.S. Census Bureau; Javier Miranda, U.S. Census Bureau; Scot Rumburg, National Agricultural Statistics Service

Key Words: Cutoff sampling, Federal surveys, Model-based estimation, Establishment surveys

Government policy makers, economic analysts, and the general public rely heavily on data gathered from federal establishment surveys for information on the U.S. economy. Faced with increasing data collection costs and concerns about heavy respondent burden, some statistical agencies -- including the Bureau of Economic Analysis, the Census Bureau, and the Energy Information Administration -- utilize "cutoff sampling," a model-based estimation technique, to maximize the information they extract from their survey data. Cutoff sampling, a highly cost-effective solution for many establishment surveys, involves selecting only the largest units in the population for the sample and using statistical models-often with auxiliary data-to extrapolate the survey information to the smaller units. In this paper, we review cutoff sampling's benefits and shortcomings, discuss experience in using cutoff sampling in federal establishment surveys, and identify situations for which cutoff sampling is well suited and others for which it is poorly suited.

New Technique for Modifying the Cutoff Sample and Its Application

✦Yang Cheng, U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20746, *yang.cheng@census.gov*

Key Words: Cutoff sampling, modified cutoff sampling, regression estimators, decision-based estimation, governments statistics

The cutoff sampling technique is used in establishment surveys in the U.S. Census Bureau. Recently, the Governments Division in the U.S. Census Bureau proposed a new method of cutoff sampling, which modified the traditional cutoff sampling techniques by constructing the cutoff points based on the size of units in the strata and then reducing the sample size in the strata with small-sized units. We also introduced a decision-based estimation method as a stratum-wise regression for strata defined first by cutoff sampling methods, and then through stratum-collapsing rules determined from the results of a hypothesis test for equality of regression slopes. Finally, we applied the modified cutoff sampling technique and decision-based estimation for two major surveys of governments: the Annual Survey of Public Employment & Payroll and the Annual Finance Survey.

Theory and Methods for Cut-Off Sampling

✦Jay Breidt, Colorado State University, CO 80523, *jbreidt@stat.colostate.edu*; Daniel Bonnery, CREST-ENSAI; Francois Coquet, CREST-ENSAI

Key Words: informative design, superpopulation, distrbution function, complex survey

Cut-off sampling is one example of informative selection from a finite population. If responses are realized as independent and identically distributed (iid) random variables from a superpopulation probability density function (pdf) f, then the joint distribution of the sample responses,given that they were selected, is not iid f. In general, the informative selection mechanism may induce dependence among the selected observations. The impact of such dependence on the empirical cumulative distribution function (cdf) is studied. An asymptotic framework and weak conditions on the informative selection mechanism are developed under which the (unweighted) empirical cdf converges uniformly, in L2 and almost surely, to a weighted version of the superpopulation cdf. We study implications of this result for estimation methodology under cut-off sampling and other informative sampling designs.

Cut-Off Approach To The Design Of Longitudinal Business Surveys

Marco Bee, University of Trento; ✦Roberto Benedetti, University of Chieti-Pescara, Viale Pindaro 42, 65127, Pescara, Italy, *benedett@unich.it*; Federica Piersimoni, Italian National Statistical Institute; Giuseppe Espa, University of Treno

Cut-off sampling is a procedure commonly used by national statistical institutes to select samples. The population is partitioned in two or three strata such that the units in each stratum are treated differently. In particular, part of the target population is usually excluded a priori from sample selection. As usual in business surveys, we assume the population of interest to be positively skewed, because of the presence of few "large" units and many "small" units. If one is interested in estimating the total of the population, a considerable percentage of the observations gives a negligible contribution to the total; on the other hand, the inclusion in the sample of the largest observations is essentially mandatory. In such situations, practitioners often use partitions of the population in three sets: a take-all stratum whose units are surveyed entirely, a take-some stratum from that a simple random sampling is drawn and a take-nothing stratum whose units are discarded. Given that practitioners are in favor of such partitions of the population and there are technical reasons that justify their use, the basic question is: is it possible to consider cut-off sampling as a valid sampling scheme? If the answer is positive, the issue is to define a statistical framework for cut-off sampling.

211 Emerging Challenges in Studies of Gene-Gene and Gene-Environment Interaction ●

ENAR, Section on Statistics in Epidemiology
Monday, August 1, 2:00 p.m.–3:50 p.m.

Results on Confounding and Misclassification for Gene-Gene and Gene-Environment Interactions

✦Tyler J VanderWeele, Harvard University, 677 Huntington Avenue, Boston, MA 02115 USA, *tvanderw@hsph.harvard.edu*

Key Words: Causal inference, Confounding, Misclassification, Gene-gene, Gene-environment

Studies of gene-gene and gene-environment interaction may be subject to confounding of either the genetic or environment factors; both factors may also be subject to misclassification. Bias formulas for sensitivity analysis for interaction under unmeasured confounding are given on both additive and multiplicative scales. Simplified formulas are provided in the case in which one of the two factors does not interact with the unmeasured confounder in its effects on the outcome. An interesting consequence of the results are that if the two exposures of interest are independent (e.g. gene-environment independence) then even under unmeasured confounding if the estimated interaction is non-zero then either there is a true interaction between the two factors or there is an interaction between one of the factors and the unmeasured confounder; an interaction must be present in either scenario. It is moreover shown that when one or both exposures are misclassified, a non-zero interaction for the misclassified exposures will imply non-zero interaction for the true exposures provided misclassification is independent and nondifferential and the sensitivity and specificity satisfy certain bounds.

Interaction Models for Longitudinal Studies with Large-Scale Genetic Data

✦Bhramar Mukherjee, University of Michigan, 1420 Washington Hgts, Ann Arbor, MI 48109, *bhramar@umich.edu*

Key Words: Singular Value Decomposition, Factor Analysis, Residual Interaction, Additive Models

There has been a large volume of literature on modeling gene-gene and gene-environment interaction in case-control studies of gene-disease association. Many longitudinal cohort studies with repeated measures on quantitative traits have collected genetic and epigenetic data in more

recent years. Limited literature is available specifically on modeling interactions in such cohort studies. In this talk I will discuss sparse modeling of interaction effects in longitudinal studies by a factor analytic representation of the interaction matrix after fitting other additive terms of the model. The idea is derived from classical psychology and design of experiments literature appearing in the 1970-1980's. We will then extend this idea to accommodate flexible time varying interaction terms in generalized additive mixed models. The methods will be illustrated by using data from the Normative Aging Study, a longitudinal cohort study of Boston area veterans.

Testing Gene-by-Environment Interactions in Behavior Genetic Designs

✦Paul Rathouz, University of Wisconsin, Dept of Biostatistics and Medical Informatics, K6/446 CSC, Box 4675; 600 Highland Ave, Madison, WI 53792, *rathouz@biostat.wisc.edu*; Carol Van Hulle, University of Wisconsin; Qianying Liu, University of Chicago

Investigators in the field of behavior genetics exploit twin and other family designs in order to decompose phenotypic variance into components that are due to aggregate genetic factors and those that are due to aggregate environmental factors. Multivariate methods have allowed for the examination of the degree of overlap in genetic and/or environmental contributions between two or more phenotypes. Newer nonlinear structural models have attempted to extend these models in order to quantify gene-by-environmental (GxE) interactions wherein the focus is on specific environmental or pre-disposing factors, and their potentially synergistic effects with aggregate genetic effects. Motivated by problems in developmental psychopathology, we critically examine the statistical performance of these GxE methods with respect to model identifiability and power. We also propose a graphical decomposition of variance that can be useful to display fitted model results.

Efficient Study Designs for Assessing Gene-Gene and Gene-Environment Interactions

✦Jinbo Chen, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 3620 Hamilton Walk, 612, Blockley Hall, Philadelphia, PA 19104 USA, *jinboche@mail.med.upenn.edu*

Using the two-phase design framework, we consider efficient study designs for assessing gene-environment interactions that allow investigators to fully exploit available genetic and environmental risk factor information. With a binary exposure and case-control study design, we derive close-form formulas for the odds ratio estimates for genetic and environmental effects under a multiplicative model for the genetic effect. We assess the power for jointly testing genetic effects and gene-environmental interactions when gene-environment independence and HWE are exploited to increase the power for assessing gene-environment interactions, particularly when external data on genotype frequencies and environmental exposure prevalence are incorporated.

212 Making Sense of Multiple Outcomes: Challenges for Randomized Trials and Observational Studies ■

Section on Health Policy Statistics, Biometrics Section, Biopharmaceutical Section, ENAR, Section on Statistics in Epidemiology, Section on Quality and Productivity, International Indian Statistical Association, Social Statistics Section

Monday, August 1, 2:00 p.m.–3:50 p.m.

The Use and Abuse of Multiple Outcomes in Randomized Trials

✦Nicholas Jon Horton, Smith College, Clark Science Center, 44 College Lane, Northampton, MA 01063-0001 USA, *nhorton@smith.edu*; Kristin M-B Tyler, Smith College; Frank Yoon, Harvard Medical School; Sharon-Lise Theresa Normand, Harvard Medical School

The inclusion of multiple outcomes to assess health interventions is a common strategy in randomized trials, as reporting just a single measure may not sufficiently characterize the effect of a treatment on a broad set of domains. While it is common practice to collect, analyze and report multiple measures, the efficient and appropriate analysis of multiple outcomes is not fully established. A number of approaches to accounting for multiple outcomes have been proposed, including multiplicity adjustments, use of composite outcomes and joint testing. A study of the use of multiple outcomes in recently published randomized trials with depression outcomes was undertaken, with some troubling findings and implications for statisticians and subject-matter researchers. Many studies reported more than one outcome, almost all of which provided no principled adjustment for multiple comparisons. Such studies may be difficult to interpret, as they have the potential for invalid conclusions due to Type I error rates. In this talk, we will describe the findings of the study, review principled approaches to multiple endpoints, and offer suggestions for researchers and regulators.

Multiplicity Adjustments for Correlated Binary Outcomes

✦Andrew C Leon, Weill Cornell Medical College, PWC Box 140, 525 East 68th Street, NYC, NY 10065, *acleon@med.cornell.edu*

Key Words: multiplicity adjustment, type I error, correlated endpoints

Clinical trials of psychopharmacologic and psychotherapeutic interventions tend to include multiple outcomes. For example, ratings of illness severity, response status, remission status, and functional impairment are commonly used in trials of treatments for bipolar disorder. The FDA recognizes that, at times, there is a need for multiple primaries. However, they make clear that in such a case multiplicity adjustments must be proposed in the protocol. That is not necessarily the case for NIMH funded protocols. Clinical investigators tend to resist multiplicity adjustments because of concern for statistical power. However, false positive treatments do not serve the clinical community. There are many approaches to multiplicity. Most disregard the correlation among outcomes. This can result in a conservative hypothesis testing strategy. The James procedure (1991) is an alternative that accounts for multiplicity among correlated binary endpoints. A simulation study compares this approach with 3 alternatives: Bonferroni, Hochberg and

unadjusted. Familywise type I error and statistical power are examined. The approaches are applied to a study of psychiatric interventions.

Joint Models and Tests of Multiple Noncommensurate Outcomes

✦Frank Yoon, Harvard Medical School, Dept. of Health Care Policy, 180 Longwood Avenue, B, MA 02115, *yoon@hcp.med.harvard.edu*; STUART Lipsitz, Division of General Medicine, Brigham and Women's Hospital, Boston, MA; Garrett A Fitzmaurice, Harvard School of Public Health; Nicholas Jon Horton, Smith College; Sharon-Lise Theresa Normand, Harvard Medical School

Key Words: Multiple outcomes, Clinical trial, Observational study, Testing, Likelihood, Quasi-likelihood

Multiple outcomes in randomized and observational studies in psychiatry are often non-commensurate, for example, measured on different scales or constructs. Standard multiplicity adjustments can control for Type I error, though such procedures can be overly conservative when the outcomes are highly correlated. Recent literature demonstrates that joint tests can capitalize on the correlation among the outcomes and are more powerful than univariate procedures using Bonferroni adjustments. However, joint tests are little used in practice, perhaps, due in part, to the specification of a joint model for the non-commensurate outcomes. Additionally, software routines to estimate joint models have not been widely publicized despite their wide availability. This work presents an evaluation of likelihood and quasi-likelihood methods for jointly testing treatment effects in a simulation study. Applications to a clinical trial and an observational study of mental health care illustrate their benefits. Adoption of these methods will lead to more efficient psychiatric clinical trials.

213 High Dimensional Inference

IMS, International Chinese Statistical Association, International Indian Statistical Association, Section on Statistical Computing

Monday, August 1, 2:00 p.m.–3:50 p.m.

Estimation of Functionals of Large Covariance Matrices

✦Harrison Zhou, Yale, , *Huibin.Zhou@Yale.edu*

Key Words: Covariance matrix, Functionals Estimation, Minimax

In this talk, we will discuss optimal estimation of various functionals of covariance matrix, and the connection to covariance matrix estimation under various norms.

Statistical Inference on Covariance Structure

✦Tony Cai, University of Pennsylvania, , *tcai@wharton.upenn.edu*

Key Words: covariance matrix, minimax rate of convergence, optimal estimation, hypothesis testing, high-dimensional inference

Covariance structure is of fundamental importance in many areas of statistical inference and a wide range of applications. In the high dimensional setting where the dimension p can be much larger than the sample size n, classical methods and results based on fixed p and large n are no longer applicable. In this talk, I will discuss some new results on optimal estimation as well as testing the structure of large covariance matrices. The results and technical analysis reveal new features that are quite different from the conventional problems.

Recovery of Sparse Signals via Conic Programming

✦Lie Wang, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 2-181, Cambridge, MA 02139, *liewang@math.mit.edu*; Alexander Belloni, Duke University; Victor Chernozhukov, Massachusetts Institute of Technology

Key Words: High-dimensional sparse model, unknown sigma, conic programming

We propose a pivotal method for estimating high-dimensional sparse linear regression models, where the overall number of regressors p is large, possibly much larger than n, but only s regressors are significant. The method is a modification of LASSO, called square-root LASSO. The method neither relies on the knowledge of the standard deviation of the regression errors nor does it need to pre-estimate. Despite not knowing the variance, square-root LASSO achieves near-oracle performance, attaining the convergence rate that matching the performance of the standard LASSO that knows the variance. Moreover, we show that these results are valid for both Gaussian and non-Gaussian errors, under some mild moment restrictions, using moderate deviation theory. Finally, we formulate the square-root LASSO as a solution to a convex conic programming problem, which allows us to use efficient computational methods, such as interior point methods, to implement the estimator.

214 Causal Diagrams and Causal Inference ●

Section on Statistics in Epidemiology, ENAR

Monday, August 1, 2:00 p.m.–3:50 p.m.

External Validity and Transportability: A Formal Approach

✦Judea Pearl, UCLA, , *judea@CS.UCLA.EDU*; Elias Bareinboim , UCLA

Key Words: causal inference, external validity, surrogate endpoint, principal strata, mediation, direct effect

We provide a formal definition for the notion of ``transportability,' or ``external validity,' which we view as a license to transfer causal information learned in experimental studies to a different environment, in which only observational studies can be conducted. We introduce a formal representation called ``selection diagrams' for expressing knowledge about differences and commonalities between populations of interest and, using this representation, we derive procedures for deciding whether causal effects in the target environment can be inferred from experimental findings in a different environment. When the answer is affirmative, the procedures identify the set of observational studies that need be conducted to license the transport. We further provide a causally principled definition of ``surrogate endpoint' as a robust pre-

dictor of effects, and show how valid surrogates can be identified in a complex network of cause-effect relationships". For the full paper, see http://ftp.cs.ucla.edu//pub/stat_ser/r372.pdf. For background material and related topics, see http://bayes.cs.ucla.edu/csl_papers.html

Causality, Conditional Independence, and Graphical Separation in Settable Systems

✦Karim Chalak, Boston College, Dept. of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA 02467 USA, *chalak@bc.edu*; Halbert White, University of California, San Diego

Key Words: causality, conditional independence, d-separation, Reichenbach principle, settable systems

We study the connections between causal relations and conditional independence within the settable systems extension of the Pearl Causal Model. Our analysis clearly distinguishes between causal notions and probabilistic notions and does not formally rely on graphical representations. We provide definitions in terms of functional dependence for direct, indirect, and total causality as well as for indirect causality via and exclusive of a set of variables. We apply these notions to formally connect causal and probabilistic conditions for conditional dependence among random vectors in structural systems. We state and prove the conditional Reichenbach principle of common cause, obtaining the classical Reichenbach principle as a corollary. Finally, we apply our approach to study notions of graphical separation, such as d-separation and D-separation in the artificial intelligence literature.

Alternative Graphical Causal Models and the Identification of Direct Effects

✦Thomas S. Richardson, University of Washington, , *thomasr@uw.edu*; James M. Robins, Harvard School of PUblic Health

Key Words: potential outcomes, direct effects, identifiability, causal graphical model, NPSEM, mediation formula

We consider four classes of graphical causal models: the Finest Fully Randomized Causally Interpretable Structured Tree Graph (FFRCISTG) of Robins (1986), the agnostic causal model of Spirtes et al. (1993), the Non-Parametric Structural Equation Model (NPSEM) of Pearl (2000), and the Minimal Counterfactual Model (MCM). The latter is referred to as 'minimal' because it imposes the minimal counterfactual independence assumptions required to identify those causal contrasts representing the effect of an ideal intervention on any subset of the variables in the graph. The causal contrasts identi?ed by an MCM are, in general, a strict subset of those identi?ed by a NPSEM associated with the same graph. We analyze various measures of the 'direct' causal effect, focusing on the pure direct effect (PDE). We show the PDE is a parameter that may be identi?ed in a DAG viewed as a NPSEM, but not as an MCM. Though bounds may be obtained under the MCM. We discuss the methodological and philosophical implications of this result.

Graphical Identification Condition of Direct and Indirect Effects with Misclassified Intermediate Endpoints

✦Manabu Kuroki, Osaka University, , *mkuroki@sigmath.es.osaka-u.ac.jp*

Key Words: Natural direct effect, Causal diagram, Controlled direct effect

This presentation deals with the problem of evaluating direct and indirect effects of a treatment on a response in randomized clinical trials where some intermediate endpoints exist but are misclassified. We propose a graphical identification condition for direct and indirect effects, which is based on the observation of several proxy variables of the intermediate endpoints. The result provides a feasible approach to evaluate direct and indirect effects in randomized clinical trials when unmeasured intermediate endpoints exist. It also provides guidance for how to choose proxy variables when designing a study protocol.

215 Statistical Graphics in Climate Research ■

Section on Statistical Graphics, Section on Statistical Computing, Section for Statistical Programmers and Analysts, Committee of Representatives to AAAS

Monday, August 1, 2:00 p.m.–3:50 p.m.

Visualizing Climate Change Data

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

Key Words: spatiotemporal data, interactive graphics, statistical graphics, climate data, maps, visualizing uncertainty

Graphics have a prominent place in reports such as those produced periodically by the International Panel on Climate Change. This talk will pull examples from these reports. We'll discuss the message that is being communicated and how it might be read from the graphic, along with alternative ways to present the data to communicate the same (or different) message. We will also look at other data collected to examine climate change, using new interactive plots in the cranvas package now available in R.

Exploring and Visualizing Uncertainty in Multidimensional Spatial Random Fields

✦Reinhard Furrer, University of Zurich, International Switzerland, *reinhard.furrer@math.uzh.ch*; Stephan Sain, NCAR

Key Words: Climate, Color schemes, MCMC posteriors, Multivariate Spatial processes

A sample of a bivariate Gaussian distribution is easily represented by isolines of the (empirical) density, or by reporting the means, standard deviations and correlation. In the case of bivariate Gaussian fields finding an appropriate representation with a straightforward interpretation is much more difficult. This is, for example, relevant to climate scientists dealing with fields of gridded temperatures and precipitation changes, for which it is crucial to summarize the relationships between both mean and both (pointwise) standard deviation fields. We present a simple software tool that allows to explore and visualize such relationships. The interactive software to quickly determine regions and clusters with similar properties. The basic idea is to use color schemes to represent several dimensions instead of a "one-dimensional" color label.

Creating Interactive Geospatial Displays of Climate Data on Google Earth

Deborah Nolan, University of California; ✦Duncan Temple Lang, University of California, Department of Statistics, 4118 Mathematical Sciences Building, Davis, CA 95616, *duncan@wald.ucdavis.edu*

Key Words: graphics, Google Earth, climate, RKML, computing, information visualization

Virtual earth browsers such as Google Earth offer an alternative medium for presenting spatial-temporal data that breaks from traditional static plots and has the potential to transform geographical information visualization. The interface makes it easy for the user to interact with a visualization and to bring in additional auxiliary information, such as geographic features, political boundaries, and others' data. With its built-in controls to layer, annotate, and animate information, virtual-earth browsers present an exciting opportunity for reaching broad audiences. We describe how to use RKML, an R package, to create Google Earth displays of climate data. We present three models for creating these displays: one is analogous to the traditional approach used in R to add points to a data region; another extends the formula language syntax in R to encompass geo-spatial and temporal concepts; and the

Interactive Graphics for Visualizing Uncertainty in Climate Model Ensembles

✦Tamara Greasby, National Center for Atmospheric Research, , *tgreasby@ucar.edu*; Stephan Sain, NCAR

Key Words: graphics, R, climate, SVG, visualization

Interactive graphics can be useful in all stages of a statistical analysis. Relationships between variables can be understood during exploratory data analysis. Model performances and residuals can be analyzed during the model fitting process. Finally results can be shared in an understandable way with collaborators from other fields. Using climate applications as a motivator, this talk will explore several basic interactive graphic techniques in R using the SVGAnnotation package (Temple Lang and Nolan, 2009). SVG graphics are a special case of the XML language and can be readily incorporated into webpages. With the addition of several JavaScript functions, a large variety of interactivity is available. This talk will show an example of an SVG graphic created with a combination of low, medium, and high level R functions. It will include features such as tool tips, radio buttons, and mouse-over functions.

Visualizing High-Resolution Climate Simulation Data Sets

✦John Clyne, NCAR, PO Box 300, Boulder, 80307 USA, *clyne@ucar.edu*

Key Words: visualization, climate

Unlike weather modeling numerical simulations of the earth's climate span not just days, but decades. Hence, climate modeling codes are capable of generating torrents of gridded data that must be subsequently analyzed with a variety of tools and techniques. As a consequence of these long time evolutions and limited computing resources, ironically, the spatial resolutions of solution grids employed in production climate modeling codes are quite coarse: a region the size of the United States is covered by only a handful of points, substantially blurring significant geographic features such as the Rocky Mountains. This lack of spatial resolution has had at least one benefit for climate researchers: graphical tools for displaying climate data can be quite simple. The demand for better climate prediction capabilities combined with the emergence of petascale supercomputing resources will change current practices. Next generation climate code will employ higher resolution grids and demand more sophisticated tools for analysis. This talk will explore state-of-the art visualization tools and methods, and their application to high-resolution climate data sets.

216 Adaptive designs one year after the draft FDA guidance - Where are we now?

■●

Biopharmaceutical Section, International Chinese Statistical Association, Scientific and Public Affairs Advisory Committee

Monday, August 1, 2:00 p.m.–3:50 p.m.

Adaptive Designs: An Industry Perspective

✦Jose Carlos Pinheiro, Johnson & Johnson PRD, 920 Rt 202 S, P.O. Box 300, Raritan, NJ 08869, *jpinhei1@its.jnj.com*

Key Words: Response-adaptive, flexible designs, simulation, modeling, dose finding

The release of the FDA draft guidance on adaptive designs has provided long awaited clarification on the Agency's position with regard to the use of these methods in clinical drug development programs. The encouraging tone of the draft guidance, particularly with regard to exploratory clinical studies, has led to renewed interest across the pharmaceutical industry in the use of response-adaptive approaches for improving the information efficiency of drug development. Considerable interest is currently focused on the quantification of the potential benefit of adaptive and other innovative approaches in comparison to more traditional strategies. Such evaluations are generally simulation-based and take into account not only the statistical properties of the approach (e.g., probability of success), but also its operational characteristics and costs (e.g., drug supply, duration). This talk will discuss key concepts related to the quantitative evaluation of alternative clinical development approaches, using real trial examples to illustrate and motivate the ideas and methods.

New Methodologies and Examples for Confirmatory Adaptive Designs

✦Werner Brannath, University of Bremen, Achterstr. 30, Bremen, 28359 Germany, *brannath@math.uni-bremen.de*

Key Words: adaptive design, flexible design, treatment selection, multiple type I error rate, response adaptive allocation

The talk will give an introduction to recently developed statistical methodologies for confirmatory adaptive designs.Such designs involve features like sample size re-estimation, treatment or subgroup selection at an interim analysis. One new topic will be a statistical methodolgy to control type I error rates in two-stage designs with response adap-

tive treatment allocation in the first part at which end treamtents are selected, and a second stage where the selected treatments are further investigated in a classical parallel group design.

Utility of Adaptive Design Software in Practical Implementation

✦Gernot Wassmer, ADDPLAN GmbH, Robert Perthel Str 77a, Koeln, 50739 Germany, *wassmer@addplan.com*

Key Words: Adaptive designs, Group sequential test, Software

Adaptive confirmatory designs refer to statistical designs that allow data-driven design changes whilst controlling the overall Type I error rate. As indicated in the draft guidance on "Adaptive Design Clinical Trials for Drugs and Biologics" the most important applications of these designs are sample size reassessment procedures, treatment arm selection procedures, and patient enrichment designs. Essentially, they can be regarded as a generalization of group sequential designs where data-driven design changes and more general applications are not foreseen. In this talk we review software products that allow the performance of confirmatory adaptive designs in terms of designing them, simulate their operation characteristics, and doing concrete analyses. Particularly, we discuss their applicability to specific adaptive designs requirements. This will extend the overview supplied in Wassmer and Vandemeulebroecke (Biometrical Journal, 2006) taking into account the current rapid development of methodological research in this area and the corresponding need for appropriate software.

217 In Memory of Professor Arnold Zellner: A review of his life and works ●

Memorial

Monday, August 1, 2:00 p.m.–3:50 p.m.

The Profound Impact of Arnold Zellner on Objective Bayesian Analysis and Foundations

✦James Berger, Duke University, , *berger@stat.duke.edu*

Key Words: Objective Bayes, Zellner

"I believe that there was a renewal of interest in Bayes's theorem in the 20th century because Jeffreys and many others adopted a pragmatic approach in deciding to work out the main estimation, testing, prediction and other problems that arise in scientific work using Bayesian and non-Bayesian methods and to compare the results. In general, Bayesian methods have won these "horse races" by providing excellent finite sample and asymptotic results, helped along by use of powerful computer techniques." This quote from Arnold Zellner tells a great deal about how and why he embraced Bayesian statistics as his primary statistical tool in scientific work. In this talk we will review the many fundamental contributions that Zellner made to objective Bayesian analysis and statistical foundations.

Arnold Zellner's Contributions to the Structural Econometric Modeling and Time Series Approach (SEMTSA)

✦Franz C. Palm, Maastricht University, , *f.palm@maastrichtuniversity.nl*

Key Words: structural modeling, time series analysis, macroeeconomic forecasting, volatility modeling, common features

In this talk I shall pay tribute to Arnold Zellner for his significant and path-breaking contributions to econometric modeling by reviewing the SEMTSA approach that he put forward almost forty years ago and to which I also had the privilege and pleasure to contribute to by collaborating closely with Arnold. Over the years Arnold continued to develop this approach and applied it to macroeconomic forecasting and modeling. A major feature of this approach is the emphasis on the sequential nature of econometric modeling, whereby coherence requirements are checked such as: the statistical properties of the marginal and conditional pdfs implied by an entertained multivariate model should be in agreement with those of the marginal and conditional pdfs as obtained from empirical analysis of the time series to be modeled. To illustrate the significance of Arnold's contributions, I shall also relate them to recent contributions to SEMTSA, for instance applying it in multivariate volatility modeling and to models which allow for comovements and other common features of the data.

Personal Recollections of the Life and Work of Arnold Zellner

✦George Tiao, Booth School of Business, University of Chicago, Chicago, IL , *George.Tiao@chicagobooth.edu*

Key Words: Bayesian statistics, seasonal adjustment, econometrics

In this talk, I will share highlights and personal recollections of Prof. Arnold Zellner's life and work in several areas. First, I will discuss his pivotal role in the development of Bayesian statistics during his early days at the University of Wisconsin. Second, I will touch on his seasonal adjustment work with the US Census Bureau in the 1970s. Finally, I will talk about his influence on the statistics and econometrics programs at the University of Chicago beginning in the early 1980s.

218 What Would Deming Have Said about Value-Added? ■

General Methodology, Section on Quality and Productivity

Monday, August 1, 2:00 p.m.–3:50 p.m.

The Multivariate Differential Effects Value-Added Model

✦Robert H. Meyer, University of Wisconsin-Madison, Value-Added Research Center, 1025 West Johnson Street, , *RHMeyer@wisc.edu*

A value-added model (VAM) is a quasi-experimental statistical model that yields estimates of the contribution of schools, classrooms, teachers, or other educational units to student achievement, controlling for non-school sources of student achievement growth, including prior student achievement, measured student and family characteristics, and

(typically) latent student growth trajectories. The objective is to facilitate valid and fair comparisons of productivity with respect to student outcomes, given that schools may serve very different student populations. The conventional value-added model imposes the restriction that a high-performing educational unit is identically high performing for all types of students, including, for example, students with low and high prior achievement and low and high income status. If this assumption is approximately true, classroom and teachers can validly be compared on the basis of a single performance indicator. However, this assumption might be incorrect: A given classroom or teacher could be very effective for students with low prior achievement, for example, but less so with talented and gifted (TAG) students. We present a generalized multivariate value-added model (which we refer to as a differential effects value-added model) that captures differences in value-added productivity (by student subgroups) across schools, classrooms, and teachers (and over time). Multivariate shrinkage is used to produce effect estimates with minimum mean squared error. We illustrate the utility of these models using data from Chicago, Los Angeles, Milwaukee, and New York City.

Value-Added Models, Student Sorting, and the Measurement of Teacher Quality

✦Jesse Rothstein, University of California, Berkeley, 2607 Hearst Avenue, Berkeley, CA 94720 United States, *rothstein@berkeley.edu*

Key Words: value added, teacher quality

A great deal of effort has gone into correctly specifying the educational production function for use in value added models (VAMs). But the intended purpose of VAMs is to estimate the causal effect of having a particular teacher. The assumptions under which a VAM identifies the teacher's causal effect are very strong, and depend as much on the process by which students are assigned to classrooms as on the specifics of the educational production function. Each of the common models relies on an implicit assumption about the assignment process that is falsified by panel data on actual student achievement histories - each model indicates that current teachers have "effects" on students' past test scores. Thus, VAM-based systems confound teacher effectiveness with differences in student assignments, with potentially important consequences for the identification of effective and ineffective teachers. Moreover, whatever effects VAMs identify appear to fade away extremely quickly, raising important questions about the value of measured achievement gains.

The Value Deming's Ideas Can Add to Educational Evaluation

✦Sharon Lohr, Arizona State University, School of Mathematical and Statistical Sciences, Box 871804, Tempe, AZ 85287-1804, *sharon.lohr@asu.edu*

Key Words: value-added models, quality improvement, ranking, randomized experiments, W. Edwards Deming

W. Edwards Deming often wrote about improving the quality of education, saying that there should be a system of education "in which teachers take joy in their work, free from fear in ranking." We revisit Deming's 14 points and philosophy of a "system of profound knowledge" in the context of value-added models, emphasizing measurement issues, designed experiments and randomization, and sources of variation. We then discuss the role of statisticians in educational reform.

219 Annals of Statistics Special Session

■●

IMS, International Indian Statistical Association

Monday, August 1, 2:00 p.m.–3:50 p.m.

Estimation of High-Dimensional Low-Rank Matrices

✦Alexandre Tsybakov, Laboratoire de Statistique, CREST, , *alexandre.tsybakov@upmc.fr*

Key Words: high-dimensional statistics, sparsity, matrix completion, optimal rates of convergence, low rank matrix estimation

This talk considers the model of trace regression, in which one observes linear combinations of entries of an unknown matrix corrupted by noise. We are particularly interested in high-dimensional setting where the dimension of the matrix can be much larger than the sample size. This talk discusses the estimation of the underlying matrix under the assumption that it has low rank, with a particular emphasis on noisy matrix completion. We consider several estimators, we derive non-asymptotic upper bounds for their prediction and estimation risks, and we show their optimality in a minimax sense on different subclasses of matrices satisfying the low rank assumption.

Estimation of High-Dimensional Matrices: Nuclear Norm Regularization and Restricted Strong Convexity

✦Martin Wainwright, Department of Statistics, University of California Berkeley, 421 Evans Hall #3860, Berkeley, CA , *wainwrig@stat.berkeley.edu*; Sahand Negahban, UC Berkeley; Alekh Agarwal, UC Berkeley

Problems that involve estimating a high-dimensional matrix that is low-rank, or well-approximated by a low rank matrix, arise in various applications. Examples include multi-task regression, identification of vector autogressive processes, compressed sensing, and matrix completion. A natural M-estimator is based on minimizing a loss function combined with a nuclear norm regularizer, corresponding to the sum of the singular values. For such estimators, we provide non-asymptotic bounds on the Frobenius norm error that hold for a general class of noisy observation models. These results are based on the loss function satisfying a form of restricted strong convexity (RSC). We show that suitable forms of this RSC condition are satisfied for many statistical models under high-dimensional scaling, including the problem of weighted matrix completion, for which restricted isometry conditions are violated. We also discuss how this same condition can be used to prove globally geometric convergence of simple iterative algorithms for solving the convex program. Based on joint works with Alekh Agarwal and Sahand Negahban.

220 Statistical Consulting in the Application of Advanced Biotechnologies to Medical Health Care ■

Section on Statistical Consulting, Section for Statistical Programmers and Analysts

Monday, August 1, 2:00 p.m.–3:50 p.m.

Annotation: Biological Meaning From Statistical Genomics

✦David Rocke, Division of Biostatistics, University of California, Davis, University of California, Davis, Davis, CA 95616, *dmrocke@ucdavis.edu*

Key Words: consulting, annotation, genomics, biology

The result of the statistical analysis of gene expression data, proteomics data, or metabolomics data is often a list of gene, proteins, or metabolites that differ significantly. These lists can be difficult for biologists to interpret, and are insufficient for publication of results in major journals. There are now many tools that can be used to explore the biological properties of lists of molecules, and for effective collaboration, statisticians need to be able to use these tools and to understand the results biologically and statistically. In this presentation, I discuss several of these tools and how to use them most effectively in statistical consulting. Illustrations are drawn from radiation biology and cancer therapy.

Tips And Traps In Using A Microarray Based Diagnostic

✦Jim Veitch, Veracyte, Inc., 7000 Shoreline Court, Suite 250, South San Francisco, CA 94080, *jveitchv@gmail.com*

Key Words: consulting, microarray, V&V

As a statistician who is a newcomer to the world of diagnostics this talk draws on my experience in working with the data analysis group in taking Veracyte's microarray based diagnostic for thyroid cancer through FDA V&V. Topics include how a discovery microarray differs from a production microarray, choosing the right metrics to measure performance, and how we handled the prescriptive nature of V&V experimental protocols.

Notes On Algorithms For Detection Of Amplicon Variants In Next-Generation Sequencing

✦Wei-min Liu, Roche Molecular Systems, Inc., 4300 Hacienda Drive, Pleasanton, CA 94588, *wei-min.liu@roche.com*; Yan Li, Roche Molecular Systems, Inc.; Julie Tsai, Roche Molecular Systems, Inc.; Mari Christensen, Roche Molecular Systems, Inc.; Wei Wen, Roche Molecular Systems, Inc.

Key Words: consulting, sequencing, algorithm, amplicon, mutation, semi-global alignment

We remark on several issues in algorithm development for detection of amplicon variants in next-generation sequencing. First, parameters of the semi-global alignment can be adjusted for various purposes. Second, compliance with the common nomenclature of mutations is helpful. Third, we propose the concept of simple variants and explain how they help define unique simple or complex variants. In addition, we propose and utilize special file formats that can significantly save storage space and reduce RAM usage.

Statistical Pharmacogenomics Studies In Data Services Cro And Biotech

✦Kit Lau, i3 Statprobe, 39120 Argonaut Way #568, Fremont, CA 94538, *Kit.Lau@i3statprobe.com*

Key Words: consulting, service, pharmacogenomics, CRO, biotech

Pharmacogenomics (PGx) studies to elucidate biomarkers to predict therapy response and safety are becoming an integral part of clinical trials. The potential of having genomic biomarkers to help tailor therapies for individuals and to predict prognosis is being realized. However, significant challenges in study design and statistical analyses in genomic data remain important road blocks to the success in the field. Statistical learning, data mining methods together with classical statistical techniques to analyze genomic data for clinical utilities would be illustrated, using examples of genetic and genomic studies to elucidate genetic association with drug response and prognostic signature in Neurosciences and Oncology. Through these examples, triumphs and tribulations of working as a CRO statistician in a strategic alliance with big pharmaceutical companies would also be discussed.

Use Of Next-Gen Sequencing To Identify Gene Mutations Beyond Kras For Evaluation Of Response To Pmab In A Randomized Phase 3 Monotherapy Study Of Metastatic Colorectal Cancer

✦Jing Huang, Amgen, Inc., 1120 Veterans blvd, ASF3 - 3013, South San Francisco, CA 94080, *jing_huang@comcast.net*

Key Words: consulting, sequencing, gene mutation, cancer, KRAS, Pmab

We used 454 sequencing to investigate whether genes beyond KRAS is predictive of response to Pmab. 288 tumor samples, balanced between two treatment arms, yielded results for a mean of 7.85 genes per patient. Mutations were detected in: 0.4% AKT1, 7.4% BRAF, 2.0% CTNNB1, 1.1% EGFR, 2.5% KRAS (exon 3), 5% NRAS, 9.4% PIK3CA, 5.5% PTEN, and 60.3% TP53. A higher than expected rate of simultaneous mutation at KRAS and either BRAF or NRAS was observed. As expected, Pmab significantly improved PFS in KRAS wild type patients and had no effect in KRAS mutant patients. In addition, mutations in NRAS were associated with lack of response to Pmab. In conclusion, 454 sequencing used to assess multiple gene loci in archival tumor samples. The superior sensitivity of 454 sequencing revealed unexpected genotypic complexity. Detailed implications of tumor genotype at these 9 loci will be presented.

221 Survey Redesign -- More than Just Questionnaire Redesign ■

Section on Government Statistics, Section on Statistical Consulting, Social Statistics Section

Monday, August 1, 2:00 p.m.–3:50 p.m.

Expanding The Frame For Nsf'S Higher Education R&D Survey: Methodological Challenges And Lessons Learned

✦Ronda Britt, National Science Foundation, 4201 Wilson Blvd. Suite 965, Arlington, VA 22230, *rbritt@nsf.gov*; Jeri Metzger Mulrow, National Science Foundation

Key Words: establishment survey, screener survey

The National Science Foundation (NSF) conducts an annual survey of roughly 800 research-performing universities and colleges called the Higher Education R&D (HERD) Survey. The survey collects data on academic R&D expenditures at 4-year and above postsecondary institutions by source of funding and field. In the past, the survey population was identified through administrative records of federal funding to universities. In 2010, NSF decided to undertake a complete census of institutions to expand the frame to include all R&D performers regardless of federal funding. A screening survey was administered to over 2000 colleges and universities to collected limited data on R&D. The major challenge of the screening effort was identifying and securing cooperation from the appropriate office within the institution. Many were not research-oriented or familiar with NSF and its mission. The mailing materials were carefully crafted to simplify the purpose of the survey and the concept of R&D. This presentation summarizes the methodology used to contact the institutions, results, and lessons learned.

The Role Of A Survey Coordinator In Establishment Surveys

✦Jeri Metzger Mulrow, National Science Foundation, 4201 Wilson Blvd. Suite 965, Arlington, VA 22230, *jmulrow@nsf.gov*

Key Words: establishment survey, survey respondent

Many establishment surveys collect data covering a variety of concepts, such as financial information, human capital, organizational structure, or specialized projects. Data on these different topics are often kept in different administrative systems within the organization. Often times, no one person in an organization is the most knowledgeable about all of the different data or has access to all of the different systems. The survey respondent often has to collect and coordinate data from a variety of different internal contacts. Several National Science Foundation surveys encounter this situation. This paper describes these surveys and how they have explicitly given the survey respondent the role of survey coordinator.

Characteristics Of Large Company Respondents To An Economic Survey

✦Stephen Keller, U.S. Census Bureau, , *Stephen.F.Keller@census.gov*; Richard S. Hough, U.S. Census Bureau; Thomas Falconer, U.S. Census Bureau; Kayla Curcio, U.S. Census Bureau

Key Words: establishment survey, data quality

The Business R&D and Innovation Survey (BRDIS) is a new survey replacing the Survey of Industrial Research and Development(SIRD). The BRDIS was fielded for the first time in 2008. It is conducted under a joint partnership agreement between the National Science Foundation and the U.S. Census Bureau. The survey covers R&D expense, R&D costs funded by others, R&D employment, innovative activities by companies, patents and intellectual property protections. Data are collected from 40,000 private sector companies in the U.S. The top 500 R&D performing companies account for more than 80% of the total R&D costs in the U.S. It is extremely important to obtain a high response rate with this set of respondents. Thus, a special communication plan was implemented in 2008 for these companies. The plan includes assigning each company to an account manager, one of the survey's analysts at the U.S. Census Bureau's headquarters, who is responsible for a series of communications with the company to assist them. As a result, we have identified several distinct characteristics in these respondents. This paper will examine the impact of these characteristics on survey response and data quality.

Fighting Confirmation Bias In Economic Surveys

✦Brandon Shackelford, Twin Ravins Consulting, 36 Lovegrass Lane, Austin, TX 78745, *brandon@twinravensconsulting.com*; Susan Helig, U.S. Census Bureau

Key Words: establishment survey, confirmation bias, redesign

This paper describes how the instructions and questions on the 2008 Business R&D and Innovation Survey (BRDIS) fell prey to confirmation bias and the steps taken to address the problem in the 2009 survey. Confirmation bias is a tendency to interpret new information in a way that confirms one's preconceptions and to ignore information and interpretations which contradict prior beliefs. Most economic surveys collect data on concepts that are commonly understood. Researchers can assume relatively little error is introduced by respondents interpreting these concepts inconsistently. During the pretesting and collection of the 2008 BRDIS evidence arose indicating that respondents interpreted the key concept of the survey inconsistently, and that the instructions were often ineffective at eliciting responses from respondents in line with the survey's definitions. To correct this problem, the 2009 BRDIS design used questions rather than instructions to guide respondents to the survey's key concept. Results indicate that the design change was somewhat effective, but that some respondents persist in reporting based on a conception that does not match the survey's definition. chang

Imputation For Nonmonotone Past-Value-Dependent Nonresponse In Longitudinal Studies With Application To The Survey Of Industrial Research And Development

✦Martin Klein, U.S. Census Bureau, , *martin.klein@census.gov*; Jun Shao, University of Wisconsin

Key Words: Bootstrap, Imputation model, Intermittent missing, Kernel Regression, Linear regression, Missing not at random

We present an overview of new regression based imputation methods for longitudinal studies with nonmonotone past-value-dependent nonrespondents. In the case of nonmonotone nonresponse, the past-value-dependent nonresponse mechanism is nonignorable as defined by Little and Rubin (2002). The methods do not require any parametric model on the joint distribution of the study variables across time points or on the response mechanism. We explore the application and customization of these methods to the Survey of Industrial Research and Development (SIRD), a survey conducted jointly by the U.S. Census Bureau and the U.S. National Science Foundation (NSF). In the

current imputation procedure used for the SIRD, total spending on research and development for a nonresponding company is imputed by the company's data from a previous year, after making an adjustment for industry growth. The proposed methods share similarities with the current procedure, while providing a framework for imputation grounded on the formal assumption of nonmonotone past-data-dependent nonresponse. We use the bootstrap to obtain variance estimates that account for uncertainty due to imputation.

222 Statistical Literacy 2011 ■●

Section on Statistical Education

Monday, August 1, 2:00 p.m.–3:50 p.m.

Pop-Stats Books And Statistical Education

✦Kaiser Fung, New York University, , *kkf2@nyu.edu*

Key Words: Popular statistics, Publishing, Curriculum, Freakonomics, Books, Education

Recently, pop-stats books have captured the public's favor, overcoming the negative perception of the subject of statistics. The best known examples include the Malcolm Gladwell series; the Freakonomics franchise; Ian Ayres's Super Crunchers; and the speaker's Numbers Rule Your World. Readers find these books highly accessible as each author finds a way to balance readability and rigor. What can educators learn from this publishing phenomenon? What is the role of pop-stats books in statistics courses? (Partly based on joint work with Andrew Gelman)

Integrating Quantitative And Financial Literacy

✦Joseph Ganem, Loyola University Maryland, Department of Physics, 4501 N. Charles Street, Baltimore, MD 21210, *comments@josephganem.com*

Key Words: quantitative Literacy, financial literacy, behavioral finance, high school math, math curriculum, instructional resources

The financial crisis of 2008-09 prompted proposals in many states to add financial literacy requirements to the high school curriculum. However, this paper proposes that financial literacy should be integrated into the current math curriculum rather than taught separately. One approach to integration is to show how quantitative thinking can be used to re-frame common financial decisions. Behavioral finance studies have shown that decision-making is highly dependent on how a financial proposition is presented or "framed." Examples are given of the use of quantitative techniques to re-frame decisions that involve buying, selling, borrowing and earning income. The implementation of Web-based instructional resources to demonstrate re-framing methods is described.The advantages to this approach are that students learn that math is relevant to personal finance and that algebra arises naturally from quantitative thinking.

Statistics And Causation: Past, Present And Future

✦Herbert Weisberg, Correlation Research, Inc., 61 Pheasant Landing, Needham, MA 02492, *hweisberg@correlation.com*

Key Words: causation, causal mechanisms, historical

Although statistical methods were originally motivated by causal ideas, overt discussion of causality had all but disappeared from statistical work by around 1900. Major statistical achievements throughout the twentieth century focused on application and elaboration of a completely non-causal paradigm. In recent decades, causality has begun to reappear in various guises, but only at the periphery of statistics. How and why did causes become so invisible to statisticians, and what are the consequences today? Can and should more explicit attention to causal mechanisms be brought into statistical methodology, teaching and practice? The answers to these questions will have a critical bearing on whether the field of statistics can maintain its relevance, and can have broad appeal to any but the most mathematically gifted students. Particularly in the biomedical and social sciences, causal analysis is much needed. These issues will be illustrated by considering the current dilemma posed by recent meta-analyses on the efficacy of aspirin therapy for prevention of heart attacks.

Learning To Read The Numbers: Critical Literacy And Numeracy In K-8 Classrooms

✦David Whitin, Wayne State University, 19836 East Ida Lane, grosse pointe woods, MI 48236, *an7657@wayne.edu*; Phyllis Whitin, Wayne State University

Key Words: critical reading, literacy and numeracy, elementary education

Being a critical reader of data is an integral part of literacy in today's information age. Developing the skills and attitudes necessary for this critical stance must begin in elementary school. To frame this instruction, the presenters developed a heuristic that provides questions to guide learners in interrogating data-related texts. Examples from primary and middle grades illustrate aspects of this heuristic, e.g. critiquing definitions, categories and visual representations. The stories show the importance of supporting children to be both critical composers and critical readers of texts. An interdisciplinary unit of study with fifth graders who examined how cereals are marketed to children further illustrates this perspective. Questions from the heuristic were used to guide students to critique their own methods of data collection and a nutritional rating system devised by Consumer Report. In this way they demonstrated a critical orientation toward data-texts over time as both readers and composers.

Assessing Quantitative Reasoning: What Do Freshmen Know?

✦Ermine Faith Orta, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, *ermine.orta@utsa.edu*; Nandini Kannan, The University of Texas at San Antonio; Kimberly Massaro, The University of Texas at San Antonio

Key Words: Assessment, Quantitative Reasoning, Instrument, Factor Analysis

The University of Texas at San Antonio (UTSA) is one of the fastest growing public universities in the state with a student population that includes a large number of historically underserved students. UTSA has developed a comprehensive plan to integrate quantitative reasoning across the core curriculum. An instrument was developed to assess the baseline quantitative literacy levels of incoming freshmen. Items on

the instrument test a student's working knowledge of simple probability, interpreting data summaries, and interpreting graphs and charts. Items on the instrument map to well-defined quantitative learning outcomes. Student performance on the instrument is grouped into three categories; at or below basic, intermediate, and proficient. Comparisons across gender and ethnicity will be discussed. In addition, performance on the assessment will be compared to student performance on SAT, ACT and Math placement tests. Results of the instrument will provide item level data which will allow for longitudinal assessment during the student's program of study. The results of a confirmatory factor analysis will be presented. In addition, the use of Rasch Models will be discussed.

223 Frontiers in Dynamic Modeling and Machine Learning

Section on Statistical Learning and Data Mining, International Society for Clinical Biostatistics, Section on Statistical Computing

Monday, August 1, 2:00 p.m.–3:50 p.m.

Random Forests

✦GÈrard Biau, UniversitÈ Pierre et Marie Curie, BoÓte 158, Tour 15-25, 2Ëme Ètage, 4 place Jussieu, Paris, 75252 France, *gerard.biau@upmc.fr*

Key Words: Random forests, Statistical learning, Sparsity, Regression estimation, Tree, Aggregation

Random forests are a scheme proposed by Leo Breiman in the 00's for building a predictor ensemble with a set of decision trees that grow in randomly selected subspaces of data. Despite growing interest and practical use, there has been little exploration of the statistical properties of random forests, and little is known about the mathematical forces driving the algorithm. In this talk, we offer an in-depth analysis of a random forests model suggested by Breiman in 04, which is very close to the original algorithm. We show in particular that the procedure is consistent and adapts to sparsity, in the sense that its rate of convergence depends only on the number of strong features and not on how many noise variables are present.

Resolving The Structure Of Interactomes With Hierarchical Agglomerative Clustering

✦Yongjin Park, Johns Hopkins University, Clark 217, Johns Hopkins Univers, 3400 N. Charles st, Baltimore, MD 21218, *ypark28@jhu.edu*

Key Words: biological networks, model comparison, clustering, link prediction, systems biology

Network clustering is a valuable approach for summarizing the structure in large networks, for predicting unobserved interactions, and for predicting functional annotations. A new algorithm, Hierarchical Agglomerative Clustering (HAC), is developed for fast clustering of heterogeneous interaction networks. This algorithm uses maximum likelihood to drive the inference of a hierarchical stochastic block model for network structure. Bayesian model selection provides a principled method both for identifying the major top-level groups and for collapsing the fine-structure of the bottom-level groups within a network. Model scores are additive over independent edge types, providing a direct route for simultaneous analysis of multiple biological interactions. In addition to inferring network structure, this algorithm generates link predictions that with cross-validation provide a quantitative assessment of performance for real-world examples. When applied to genome-scale data sets representing several organisms and interaction types, HAC provides the overall best performance in link prediction when compared with other clustering methods and with model-free graph diffusion kernels.

On The Suboptimality Of The Mle And All Other Point Estimators When The Model Is Wrong, And Their Improvement Via Flattening

✦Wojciech Kotlowski, Centrum Wiskunde & Informatica (CWI), Science Park 123, Amsterdam, 1098 XG Netherlands, *kotlowsk@cwi.nl*; Peter Gr¸nwald, Centrum Wiskunde & Informatica (CWI); Steven de Rooij, Centrum Wiskunde & Informatica (CWI)

Key Words: exponential family, KL-risk, prequential analysis, model misspecification, flattened maximum likelihood

We analyze the KL-risk of point estimators for density estimation with exponential family models. It is known that when the data are generated by one of the distributions in the model, the maximum likelihood estimator (MLE) admits the optimal rate of convergence with the optimal constant factor in front, but when the model is misspecified (data comes from a distribution outside the model), MLE behaves suboptimally. We show that in the misspecified case, not only MLE, but every point estimator is unable to achieve the optimal constant in front of the convergence rate; the additional factor can become arbitrarily large. We then provide a solution to this problem by introducing a simple ``flattening' of the ML distribution, that does achieve the optimal performance even when the model is wrong. The flattened ML distribution lies slightly outside the model under consideration, and is thus not a point estimator. We can apply the flattened ML estimator in Dawid's prequential model selection criterion. Simulations show a major improvement of the resulting model selection criterion over standard prequential model selection based on MLE.

Probability Machines

✦James D. Malley, National Institutes of Health, CIT, Bldg. 12A, Rm. 2052, Bethesda, MD 20892, *jmalley@mail.nih.gov*

Key Words: classification, pattern recognition, learning machines

Many statistical learning machines can provide an optimal classification for binary outcomes. However, probabilities are required for risk estimation using individual patient characteristics for personalized medicine. This talk shows that any statistical learning machine that is consistent for the nonparametric regression problem is also consistent for the probability estimation problem. These will be called probability machines. Probability machines discussed include classification and regression random forests and two nearest-neighbor machines, all of which use any collection of predictors with arbitrary statistical structure. Two simulated and two real data sets with binary outcomes illustrate the use of these machines for probability estimation for an individual.

Dynamic Logistic Regression And Dynamic Model Averaging For Binary Classification

✦Tyler McCormick, Columbia University, Dept of Statistics, MC 4690, 1255 Amsterdam Ave, Amsterdam Ave, NC 10027 United States, *tyler@stat.columbia.edu*; Adrian Raftery, University of Washington; David Madigan, Columbia University; Randall Burd, Children's National Medical Center

Key Words: Bayesian model averaging, Binary classification, Confidentiality, Hidden Markov model, Laparoscopic surgery, Markov chain

We propose an online binary classification procedure for cases when there is uncertainty about the model to use and when parameters within a model change over time. We account for model uncertainty through Dynamic Model Averaging (DMA), a dynamic extension of Bayesian Model Averaging (BMA) in which posterior model probabilities are also allowed to change with time. We do this by applying a state-space model to the parameters of each model and a Markov chain model to the data-generating model, allowing the "correct" model to change over time. Our method accommodates different levels of change in the data-generating mechanism by calibrating a "forgetting" factor. We propose an algorithm which adjusts the level of forgetting in a completely online fashion using the posterior predictive distribution. Our algorithm allows the model to accommodate various levels of change in the data-generating mechanism at different times. We apply our method to data from children with appendicitis who receive either a traditional (open) appendectomy or a laparoscopic procedure.

224 Customer Intelligence ■

Section on Statistics and Marketing
Monday, August 1, 2:00 p.m.–3:50 p.m.

Assessing The Added Value Of Different Data Types For Predicting Customer Attrition

✦Michel Ballings, Ghent University, Tweekerkenstraat 2, Gent, 9000 Belgium, *michel.ballings@ugent.be*; Dirk Van den Poel, Ghent University

Key Words: predictive analytics, customer intelligence, survey data, classification, data types, customer churn

In an increasingly digital world, every action a customer takes is stored in data repositories, leaving companies with an ever-growing amount of data. Although this data has already been assigned the status of a strategic asset, limited assessment has been made concerning the differential value of different data types as resources in a customer intelligence context. In this study, we first elaborate on how data is important in an analytical Customer Relationship Management context, second we provide a taxonomy of different data repositories using the customer's purchasing process as an operational yardstick and third, we empirically determine the added value of each data repository in predicting customer churn. The latter issue is investigated by comparing alternative statistical as well as data mining classification techniques. The outcomes of this study provide clear recommendations as to which data types companies should invest in for optimizing the value of their resources.

Improving Customer Relationship Management Models By Including Neighborhood Effects Using Spatial Econometrics

✦Philippe Baecke, Ghent University, Tweekerkenstraat 2, Gent, 9000 Belgium, *philippe.baecke@ugent.be*; Dirk Van den Poel, Ghent University

Key Words: predictive analytics, customer acquisition, CRM, spatial models, neighborhood effects, spatial correlation

Customer relationship management (CRM) uses data mining techniques to support decisions about several marketing strategies such as customer acquisition, development and retention. From these three applications, customer acquisition models suffer the most from the fact that only a limited amount of information is available about potential prospects resulting in a relative low accuracy of the models. This study tries to improve the predictive performance of such models by including neighborhood effects. Traditional CRM models ignore the fact that correlation, resulting from social influences and homophily, could exist between the purchasing behaviors of surrounding prospects. Hence, a spatial model is used to capture these neighborhood effects and improve the identification of prospects for automobile brands. Moreover, an optimization is presented concerning the measurement level of the neighborhoods.

Binary Quantile Regression: A Bayesian Approach Based On The Asymmetric Laplace Distribution

✦Dries F. Benoit, Ghent University, Tweekerkenstraat 2, Gent, 9000 Belgium, *dries.benoit@ugent.be*; Dirk Van den Poel, Ghent University

Key Words: quantile regression, asymmetric laplace distribution, bayesian, dichotomous response data

This paper develops a Bayesian method for quantile regression for dichotomous response data. The frequentist approach to this type of regression has proven problematic in both optimizing the objective function and making inference on the parameters. By accepting additional distributional assumptions on the error terms, the Bayesian method proposed sets the problem in a parametric framework in which these problems are avoided. To test the applicability of the method, we ran two Monte-Carlo experiments and applied it to Horowitz' (1993) often studied work-trip mode choice dataset. Compared to previous estimates for the latter dataset, the method proposed leads to a different economic interpretation.

In-Store Shopping Path And Markov Chain Model

✦Katsutoshi Yada, Kansai University, 3-3-35, Yamate, Suita, International 5648680 Japan, *yada@ipcku.kansai-u.ac.jp*; Xiao-Jun Ding, Kansai University; Asako Ohno, Shijonawate Gakuen Junior College

Key Words: shopping path, Markov chain, marketing, RFID

The purpose of this research is to propose a marketing application of Markov chain model by using shopping path data. The RFID-enabled shopping carts are being put into practical use within some Japanese supermarkets for the purpose of tracking customers' shopping footprints

and collecting shopping path data. We propose a method to analyze the customer movement in a store based on Markov chain model, so as to discovery useful knowledge for store managers and marketing stuff.

Data-Oriented Hospital Services

✦Shusaku Tsumoto, Shimane University, 89-1 Enya-cho, Izumo-city, Shimane, 693-8501 Japan, *tsumoto@computer.org*; Shoji Hirano, Shimane University; Shusaku Tsumoto, Shimane University

Key Words: customer intelligence, personalized medicine, hospital information system

About twenty years have passed since clinical information are stored electronically as a hospital information system. Stored data ranges from accounting information to laboratory data and even patient records are now starting to be accumulated. If the implementation of electronic patient records is progressed into the improvement on the efficiency of information retrieval, it may not be a dream for each patient to benefit from the personal database with all the healthcare information, from cradle to tomb. However, although the studies on electronic patient record has been progressed rapidly, reuse of the stored data has not yet been discussed in detail. In this paper, data stored in hospital were analyzed. The results show several interesting results, which suggests that the reuse the reuse of stored data will give a powerful tool to support hospital management.

225 Statistical Rigor in Test and Evaluation

Section on Statistics in Defense and National Security, International Indian Statistical Assoc., Reps. for Young Statisticians, Section on Government Statistics, Section on Physical and Engineering Sciences, Section on Quality and Productivity, Section on Risk Analysis

Monday, August 1, 2:00 p.m.–3:50 p.m.

Estimating The Probability Of Engagement Success For Missile Defense Systems

✦Carl Gaither, IDA, , *cgaither@ida.org*

Key Words: Missile Defense, Probability of Success

We present a methodology for estimating the probability of engagement success (PES) for missile defense systems specifically for the portions of their engagement battle space sampled through developmental and operational testing. This methodology is based on the concept of the DuPont model which is widely used in financial analysis. It incorporates ground- and flight-test data to the maximum extent possible. The ability to incorporate as much data as possible is crucial because the cost and logistics of performing end-to-end missile defense system tests places severe restrictions on the amount of data available. This methodology offers several distinct advantages: 1) it is straightforward and offers clear insight into the missile defense functions that drive performance; 2) it identifies over- and under- performing missile defense functions; 3) it can incorporate historical testing, component-level testing, and end-to-end testing of missile defense systems; and 4) it lends itself to adaptation as missile defense systems mature. We will present a description of this methodology as well as numerical studies of its performance for realistic test outcomes.

Design For Reliability Using Robust Parameter Design

✦Laura Freeman, Institute for Defense Analyses, , *lfreeman@ida.org*

Key Words: Design for Reliability, Design of Experiments, Robust Parameter Design

The principals of Design of Experiments (DOE) recently have been widely implemented as a method of increasing the statistical rigor of operational tests. The focus has been on ensuring coverage of the operational envelope in terms of system effectiveness. DOE is applicable in reliability as well. One area the 0009 advocates for is early in the product development using a Design for Reliability (DfR) process to design in reliability. Robust parameter design (RPD) first used by Taguchi and then by the Response Surface Community provides insights to DOE can be used to make a product/process invariant to changes in factors. Using the principles of RPD, new developments in reliability analysis and the guidance of the 0009 I propose a new application of RPD to DfR.

Test Concept Development For The Ballistic Missile Defense System

✦Jasmina Marsh, IDA, , *jmarsh@ida.org*; Dawn Loper, IDA; Carl Gaither, IDA; Michael Luhman, IDA

Key Words: Ballistic Missile Defense, Department of Defense, Scenario Based Design

The Ballistic Missile Defense System (BMDS) is a major Department of Defense acquisition program with no formal requirements to be demonstrated in testing, and no formal milestone decision points. Cost and logistical issues present significant test restrictions that often limit the amount of operationally realistic testing possible before system deployments. This talk presents a rigorous methodology for developing test concepts to characterize the effectiveness of the BMDS. The method is based on modeling and analysis of the intended operational use of the system for a variety of potential threats, producing a set of operational scenarios. Statistical techniques can then be applied to the resultant data from these scenarios to determine the range and likelihood of how a threat may present itself to the BMDS assets. Unique features of these scenarios are extracted, and the scenarios are examined for the desired relative locations of assets for testing. A set of tests are then developed that maximize the operational features while staying within the confines of the test restrictions. Notional examples of this methodology will be presented.

Statistical Rigor In Operational Test And Evaluation: Success Stories And Future Challenges

✦Catherine Warner, Director of Operational Test and Evaluation, 1700 Defense Pentagon, Washington, DC 20301-1700 USA, *Catherine.Warner@osd.mil*

Key Words: Operational Testing, Department of Defense, Design of Experiments

The Director, Operational Test and Evaluation (DOT&E) is responsible for oversight of all Operational Tests and Evaluations within the Department of Defense. An operational field test is required before a Major Defense Acquisition Program can go into full-rate production. The Science Advisor's office is responsible for advising DOT&E on current, relevant scientific best practices. DOT&E has had a long standing relationship with the National Research Council's Committee on National Statistics. Recently, DOT&E, has pushed for the implementation of many recommendations of the committee including the use of Experimental Design in constructing operational tests. This talk covers examples where experimental design has improved operational test. Additionally, I propose areas for future statistical research based on challenges my office has encountered in the implementation of design of experiments to operational testing.

226 Solving Non-Standard Problems for National Surveys ■●

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

Monday, August 1, 2:00 p.m.–3:50 p.m.

Is It Feasible To Use A Sampling List Frame To Evaluate Misclassification Errors Of An Area-Frame Based Survey?

✦Denise A Abreu, USDA/NASS, 3251 Old Lee hwy Room 305, Fairfax, VA 22030, *denise_abreu@nass.usda.gov*; Andrea C Lamas, USDA/NASS; Hailin Sang, National Institute of Statistical Sciences; Pam Arroway, North Carolina State University; Kenneth Kyle Lopiano, University of Florida; Linda J Young, University of Florida

Key Words: Misclassification Errors, Area Frame, List Frame, Record Linkage, Re-screening Survey

During the past three years, the National Agricultural Statistics Service (NASS) has made an effort to address, quantify and adjust for an undercount in the number of farms indication on its annual June Area Survey (JAS), which is based on an area frame. This undercount is a direct result of the misclassification of agricultural tracts as non-agricultural. The 2007 Census of Agriculture mailing list (CML) was evaluated as a potential source to assess misclassification on the 2007 JAS. This evaluation revealed that the CML was a rich source from which to quantify the undercount of farms on the JAS. However, the CML is only available every five years and misclassification on the JAS should be assessed each year. Independently of the area frame, NASS maintains a list of agricultural operators, referred to as the list frame. Yearly list-based samples are selected from the list frame. In addition, the list frame serves as the foundation for building the CML. The list frame is updated on an on-going basis and operators are categorized as either active or inactive. This paper discusses the feasibility of using the list frame to assess misclassification on the JAS.

Adjusting An Area Frame Estimate For Misclassification Using A List Frame

✦Andrea C Lamas, USDA/NASS, 3251 Old Lee hwy Room 305, Fairfax, VA 22030, *Andrea_Lamas@nass.usda.gov*; Denise A Abreu, USDA/NASS; Hailin Sang, National Institute of Statistical Sciences; Kenneth Kyle Lopiano, University of Florida; Pam Arroway, North Carolina State University; Linda J Young, University of Florida

Key Words: Misclassification, Area Frame, List Frame, Classification and Regression Tree, Receiver Operating Characteristic curve

In recent years, the National Agricultural Statistics Service (NASS) evaluated a variety of approaches to adjust for misclassification in its annual June Area Survey (JAS), which is based on an area frame. This misclassification is a direct cause of an undercount in the number of farms indication produced by the JAS. One approach to correct for this undercount is to use NASS's sampling list frame, which is independent of the area frame. However, recent studies showed that there are farm status inaccuracies on the list frame. These are active records that are not associated with farms. If the list frame farm status inaccuracies are not addressed, the adjusted JAS number of farms indication could become biased upwards. Using Classification and Regression Tree (CART) models, the probability that a list frame record is active can be obtained. This paper evaluates methods for classifying each active list frame record as either a farm or non-farm. One method sets a cut-off using the probabilities, a Receiver Operating Characteristic (ROC) curve, and auxiliary data. Another method uses the same probabilities along with auxiliary data in adjusting for misclassification.

Estimating Change For A Dynamic Target Population

✦Ivan A. Carrillo-Garcia, NISS, 19 T.W. Alexander Drive, Research Triangle Park, NC 27709, *ivan@niss.org*

Key Words: Longitudinal surveys, Panel surveys, Weighting, Finite population, Super-population, Division of Science Resources Statistics, NSF

When the finite population of interest is dynamic, a common way to estimate change is to fix a target population as that at one fixed time, often the first wave. However, if the population steadily accrues new individuals, this requires either disregarding some data or using fluctuating weights. If individuals also leave the actual population, complexities increase. An important case is the Survey of Doctorate Recipients, a panel survey conducted by the NSF, that collects data on doctoral holders in science, engineering, or health. The goal is to estimate socioeconomic changes of the doctorate holders and to assess these across fifteen years. The major difficulties are that the target population has changed during the period of interest, with the addition of new degree holders and with loss or deletion of other subjects. Hence the usual setting for finite population inference does not occur; rather there are a set of different (albeit overlapping) finite populations through time. A superpopulation formulation addresses these difficulties. A hypothetical model generates the different finite populations; and the quantities of interest are the parameters of this superpopulation model.

Three-Phase Sampling Design Estimation With Application To Non-Response

✦Hailin Sang, National Institute of Statistical Sciences, 19 T.W. Alexander Drive, Research Triangle Park, NC 27709, *sang@niss.org*; Kenneth Kyle Lopiano, University of Florida; Pam Arroway, North Carolina State University; Denise A Abreu, USDA/NASS; Andrea C Lamas, USDA/NASS; Linda J Young, University of Florida

Key Words: three-phase sampling design estimation, unbiased estimator, variance estimation, non-response

Each year the National Agricultural Statistics Service (NASS) publishes an estimate of the number of farms in the United States based on the June Area Survey (JAS). Independent studies showed that the JAS number of farm indications have significant undercount due to misclassification. To adjust for this undercount, a follow-on survey to the JAS called the Annual Land Utilization Survey (ALUS) was proposed. ALUS can be treated as a second phase to the JAS. The two-phase stratified design, JAS-ALUS, can be applied to any estimate produced by the JAS. However, ALUS has non-response. In this paper, a three-phase sampling design estimation is developed based on the two-phase sampling design estimation studied by Sarndal and Swensson (1987). We allow a general sampling design in each phase; that is, the inclusion probabilities in each phase are arbitrary. The estimator is unbiased and we provide an unbiased estimator for the variance. This method can be applied to a two-phase sampling design estimation with the third phase being response/non-response. Specifically, we can use this methodology for JAS-ALUS, with the third phase design addressing non-response.

Developing A Set Of Decision Rules For A Responsive Split Questionnaire Design

✦Jeffrey Gonzalez, U.S Bureau of Labor Statistics, Office of Survey Methods Research, 2 Massachusetts Ave NE, Suite 1950, Washington, DC 20212 USA, *gonzalez.jeffrey@bls.gov*

Key Words: Composite estimation, Cost-error tradeoffs, Matrix sampling, Multiphase sampling, Paradata, Propensity model

Responsive designs use data collected about the sample to inform mid-survey decisions, or actions, affecting the error properties of the resulting statistics. A key element of these designs is the decision rule, a function of the collected data to an action. Before survey practitioners can choose actions that will positively affect the statistics' error properties, they must understand the relationships in the data so that they are more informed about tradeoffs when choosing among actions. Responsive designs can be utilized in many settings, e.g., split questionnaires. A responsive split questionnaire would entail administering subsets of questions of a full questionnaire to sample members based on the data collected about the sample during an initial phase. We use data from the Consumer Expenditure Survey, a panel survey collecting information on the spending habits of consumers, to explore the extent to which information from the first interview, or initial phase, is related to events, e.g., purchases, in later interviews. With this information, we develop decision rules regarding which subsets of questions to administer to sample members at subsequent phases of data collection.

227 2011 David P. Byar Young Investigator Award

Biometrics Section

Monday, August 1, 2:00 p.m.–3:50 p.m.

A Generalized Least Squares Matrix Decomposition With An Application To Neuroimaging

✦Genevera I. Allen, Baylor College of Medicine & Rice University, Department of Statistics, 6100 Main St., MS 138, Houston, TX 77005 USA, *gallen@rice.edu*; Logan Grosenick, Stanford University; Jonathan Taylor, Stanford University

Key Words: matrix decomposition, principal components analysis, sparse principal components analysis, functional principal components anaylsis, neuroimaging, transposable data

Variables in high-dimensional data sets common in neuroimaging and genomics often exhibit complex dependencies. Conventional multivariate analysis methods often ignore these relationships, that can arise, for example, from spatio-temporal processes or network structures. We propose a generalization of the SVD that is appropriate for transposable matrix data, or data in which neither the rows nor columns can be considered independent instances. By finding the best low rank approximation of the data with respect to a transposable quadratic norm, our decomposition, entitled the Generalized least squares Matrix Decomposition (GMD), directly accounts for dependencies in the data. We also regularize the factors, introducing the Generalized Penalized Matrix Factorization (GPMF). We develop fast algorithms using the GMD to preform Generalized PCA (GPCA) and the GPMF to preform sparse GPCA and functional GPCA on massive data sets. Through simulations and a whole-brain functional MRI example we demonstrate the utility of the GMD and GPMF for dimension reduction, sparse and functional signal recovery and feature selection with high-dimensional transposable data.

Grouped Variable Selection Via Hierarchical Linear Models

✦Sihai Dave Zhao, Harvard University/Dana-Farber Cancer Institute, , *szhao@hsph.harvard.edu*; Yi Li, Dana-Farber Cancer Institute and Harvard School of Public Health

Key Words: Biological pathways, Grouped variables, Hierarchical linear model, High-dimensional data, Variable selection

Incorporating prior information into predictive models can often lead to better prediction and model selection. For example, in many genomic studies this prior information takes the form of grouped covariates, or biological pathways. Current methods can achieve variable selection at the group level as well as the within-group level. Framing these methods as hierarchical linear models, we find that they are actually suboptimal when used with the small samples, high-dimensional covariates, and large group sizes that are characteristic of modern high-throughput datasets. Motivated by a Bayes argument, we propose a new class of penalty functions that are designed to correct the deficiencies of existing techniques. We show that they can achieve the oracle

property under mild regularity conditions. Simulation studies illustrate the advantages of our proposal, which are further demonstrated by an application to a clinical study of multiple myeloma.

Measuring Reproducibility Of High-Throughput Experiments

✦Qunhua Li, Univ of California at Berkeley, Dept of Statistics, 367 Evans Hall, MS 3860, Berkeley, CA 94720-3860, *qli@stat.berkeley.edu*; James Ben Brown, Univ of California at Berkeley; Haiyan Huang, University of California at Berkeley; peter j Bickel, UC Berkeley

Key Words: reproducibility, copula, mixture model, irreproducible discovery rate, high-throughput experiment, genomics

Reproducibility is essential to reliable scientific discovery in high-throughput experiments. In this work, we propose a unified approach to measure the reproducibility of findings identified from replicate experiments and identify putative discoveries using reproducibility. Unlike the usual scalar measures of reproducibility, our approach creates a curve, which quantitatively assesses when the findings are no longer consistent across replicates. Our curve is fitted by a copula mixture model, from which we derive a quantitative reproducibility score, which we call the "irreproducible discovery rate" (IDR) analogous to the FDR. This score can be computed at each set of paired replicate ranks and permits the principled setting of thresholds both for assessing reproducibility and combining replicates. Since our approach permits an arbitrary scale for each replicate, it provides useful descriptive measures in a wide variety of situations to be explored. We study the performance of the algorithm using simulations and give a heuristic analysis of its theoretical properties. We demonstrate the effectiveness of our method in a ChIP-seq experiment.

Landmark Prediction Of Survival Incorporating Short Term Event Information

✦Layla Parast, Harvard University, 677 HuntingtonAvenue, Boston, MA 02115, *LPARAST@HSPH.HARVARD.EDU*; Su-Chun Cheng, Dana Farber Cancer Institute; Tianxi Cai, Harvard University

Key Words: survival analysis, risk prediction, time-varying coefficient model

In recent years, genetic and biological markers have been examined extensively for their potential to signal progression or risk of disease. In addition to these markers, it has often been argued that short term outcomes may be helpful in making a better prediction of disease outcomes in clinical practice. Due to the potential difference in the underlying disease process, patients who have experienced a short term event of interest may have very different long term clinical outcomes from the general patient population. Most existing methods for incorporating censored short term event information in predicting long term survival focus on modeling the disease process and are derived under parametric models in a multi-state survival setting. In this paper we propose to incorporate short term event time information up to a landmark point along with baseline covariates via a flexible varying-coefficient model. The performance of the resulting landmark prediction rule is evaluated non-parametrically and compared to prediction rules constructed using the baseline covariates only. Simulation studies and an example suggest that the proposed procedures perform well in finite samples.

Risk Classification With An Adaptive Naive Bayes Kernel Machine Model

✦Jessica Minnier, Harvard University, MA 02115 USA, *jminnier@hsph.harvard.edu*; Tianxi Cai, Harvard University; Jun Liu, Harvard University

Key Words: Risk prediction, Genetic association, Kernel PCA, Genetic pathways, Gene-set analysis, Kernel Machine Regression

The complex genetic architecture of disease makes it difficult to identify genomic markers associated with disease risk. Standard risk prediction models often rely on additive or marginal relationships of markers and the phenotype of interest. These models perform poorly when associations involve interactions and non-linear effects. We propose a multi-stage method relating markers to disease risk by first forming gene-sets based on biological criteria. With a naive bayes kernel machine model, we estimate gene-set specific risk models that relate each gene-set to the outcome. Second, we aggregate across gene-sets by adaptively estimating weights for each set. The KM framework models the potentially non-linear effects of predictors without specifying a particular functional form. Estimation and predictive accuracy are improved with kernel PCA in the first stage and adaptive regularization in the second stage to remove non-informative regions from the final model. Prediction accuracy is assessed with bias-corrected ROC curves and AUC statistics. Numerical studies suggest that the model performs well in the presence of non-informative regions and both linear and non-linear effects.

228 The use of historical data priors in the design and analysis of clinical trials

Biopharmaceutical Section, Section on Health Policy Statistics
Monday, August 1, 2:00 p.m.–3:50 p.m.

Evaluating Probability Of Study Success Through Bayesian Evidence

✦Haoda Fu, Eli Lilly and Company, Lilly Corporate Center, Drop code 2232, Indianapolis, IN 46285 USA, *fu_haoda@lilly.com*

Key Words: Bayesian Method, Probability of Study Success, BEST, Clinical Trial Optimization

To optimize drug development, we have to think in three different levels: individual study level, compound level, and portfolio level. No matter in which level, there are many choices we have to consider. The probability of study success is the key to evaluate different choices. In this paper, we propose a Bayesian method to evaluate the probability of study success (PrSS) to evaluate different study choices and thus to facilitate quality decision making. The distinction between traditional power and the PrSS is described, and the applications of PrSS will be shown by examples. The optimization in different levels will be discussed.

A Bayesian Adaptive Dose Selection Procedure With A Count Endpoint

✦Luca Pozzi, University of California Berkeley, Division of Biostatistics, 4180 Opal Street, Apt. 10, Oakland, CA 94609, *p.luc@stat.berkeley.edu*; Heinz Schmidli, Novartis Pharma

AG; Mauro Gasparini, Politecnico di Torino, Department of Mathematics; Amy Racine, Novartis AG

Key Words: Clinical Trials, Bayesian Model Averaging, Predictive Probability, Adaptive Design, Dose Selection

In clinical drug development, a sequence of studies is carried out to identify an efficacious and safe dose of a newly developed pharmaceutical drug. Adaptive designs Can considerably improve upon traditional designs, by modifying design aspects of the ongoing trial, including early stopping, adding or dropping doses, or changing the sample size. In the present work we propose a two-stage Bayesian adaptive design for a Phase II study aimed at selecting the lowest e?ective dose for Phase III. In the first stage patients are randomized to placebo, maximal tolerated dose, and one or more additional doses within the dose range. Based on an interim analysis, the study is either stopped for futility or success, or enters the second stage, where newly recruited patients are allocated to placebo, maximal tolerated dose, and one additional dose chosen based on interim data. Assuming a monotone dose-response relationship, at interim, criteria based on the Predictive Probability of Success are used to decide on whether to stop or to continue the trial, and, in the latter case, which dose to select for the second stage.

From Historical Data To Priors: Methodological And Practical Aspects

✦Beat Neuenschwander, Novartis Pharma, WSJ-103.1.26.6, Basel, 4056 Switzerland, *beat.neuenschwander@novartis.com*

Key Words: clinical trials, meta-analysis, prediction, priors

The formal use of contextual evidence beyond the actual study data is a potentially useful approach to decision making in clinical trials. However, this is easier said than done. The challenges are methodological as well as practical. This presentation will provide an overview of methodological approaches to historical data that allow for reasonable discounting of historical data due to different degrees of similarity between historical and current data. Examples from phase I, II and IV clinical trials will be used to highlight practical aspects of using historical data. The examples comprise a phase II proof-of-concept study that resulted in a conflict between historical and actual data, the design of a Japanese phase I trial based on Western data using mixture priors in order to robustify prior assumptions, and a large post-marketing phase IV study where a rich historical database was available but eventually not used. While historical data are often a useful source of information, using these data sensibly requires good judgement paired with appropriate Bayesian statistical methodology.

Examples Of Safety And Efficacy Analyses In Early Clinical Development Incorporating Historical Data

✦Neal Thomas, Pfizer Inc, 61 Dreamlake Drive, Madison, CT 06443 USA, *snthomas99@yahoo.com*

Key Words: Bayes, historical data, safety assessment, dose response

Several examples will be presented which used historical data to create Bayesian prior distributions. The examples include settings in which internal patient level data are available, and settings in which only published historical summaries are available. Some of the prior information is compound-specific, and some of it is based on broader experience with drug development. Applications include safety and efficacy assessments.

229 Novel Bayesian Methods for Biostatistics ●

Section on Bayesian Statistical Science, ENAR

Monday, August 1, 2:00 p.m.–3:50 p.m.

Bayesian Semiparametric Nonstationary Correlation Models For Longitudinal Data

✦Robert E. Weiss, University of California, Los Angeles, , *robweiss@ucla.edu*; Lei Qian, Research & Evaluation Department, Kaiser Permanente, Southern California

Key Words: Change point, CD4 Cell counts, Growth curve, B-spline, Covariance model

We develop a matrix mixture correlation model for continuous longitudinal data with nonstationary correlations. We consider two situations, one where there is a known change point causing nonstationarity, and a second situation where the correlations change smoothly over time. The former is appropriate when a known change in the biological system has a substantial impact on outcomes. The second situation with continuously changing correlation happens for example in growth curve studies where growth in early life is very different from later growth. Our model allows both the correlation values and underlying structure to change over time and is capable of handling highly unbalanced data with a large number of repeated measurements.

Application Of Bayesian Statistics To Electronic Medical Records

✦Song Zhang, University of Texas Southwestern Medical Center, Dallas, TX 75390, *song.zhang@utsouthwestern.edu*

An increasing number of hospitals have implemented electronic medical record (EMR) systems. One important challenge faced by biostatisticians is to extract meaningful information from the ever expanding EMR databases and assist medical decision making. In this study we develop statistical methods to detect signs of clinical deterioration during patients' hospitalization, which enables health care providers to direct focused attention to such patients before an adverse event occurs. One difficulty in utilizing EMR data is the frequent occurrence of missing data. A patient's health status is characterized by the current values, as well as the temporal trends, of a large number of variables. It is inevitable that at any given time point, a significant portion of variables involved in the statistical model would be missing. We propose a Bayesian hierarchical method to impute missing data in EMR. It decompose the measurement of each variable into two elements, one representing a smooth temporal trend and the other the tendency to change concurrently with other health variables. We further described how to efficiently apply the statistical learning to support real-time computation.

Modeling Hiv Behavioral Intervention Using A Hierarchical Longitudinal Nested Poisson Model

✦Yuda Zhu, UCLA, 3770 Keystone Ave, Apt 105, Los Angeles, CA 90034, *yudazhu@gmail.com*

Key Words: HIV Prevention, MCMC, Hierarchical Nested, Behavior, Multivariate Count Data

We present a multivariate nested count model to model the sexual behaviors of participants in an HIV behavioral intervention trial. The model we present is motivated by data collected from the Health Living Project (HLP) to evaluate the effects of a randomized counseling intervention. Standard models in the field typically model a single composite outcome longitudinally over time. We make the case for the inclusion of a multivariate nested outcome model. Outcomes are deaggregated on a partner specific level resulting in a 2 level nested structure. Nested within participant are the number of protected and unprotected sex acts with each individual partner. At the participant level, outcomes include the number of HIV positive and HIV negative partners. A corresponding 2 level correlation structure is used to model correlation between outcomes.

Discriminative Information Analysis In Mixture Modelling

✦Lin Lin, Department of Statistical Science, Duke University, 27708, *lin@stat.duke.edu*; Cliburn Chan, Duke University; Mike West, Duke University

Key Words: Bayesian expectation-maximization, Bayesian mixture models, Concordance of densities, Flow cytometry data, Multivariate mixture model, Subset selection

We discuss the evaluation of subsets of variables for the discriminatory evidence they provide in multivariate mixture modelling for classification. A new approach to discriminative information analysis uses a natural measure of concordance between mixture component densities. The approach is both effective and computationally attractive for routine use in assessing and prioritizing subsets of variables according to their roles in discrimination of one or more components. We relate the new discriminative information measure to misclassification rates, exemplify its use in Bayesian mixture models using novel Bayesian expectation-maximization estimation and Markov chain Monte Carlo methods, and use simulated data as an illustrative example. An application comes from the context of automatic classification and discriminatory variable selection in high-throughput systems biology using large flow cytometry data sets.

A Joint Markov Chain Model For The Association Of Two Longitudinal Binary Processes

✦Catherine M. Crespi, University of California Los Angeles, , *ccrespi@ucla.edu*; Sherry Lin, University of California Los Angeles

Key Words: Markov model, longitudinal data, joint model, nonhomogeneous

In some longitudinal studies, several related processes are measured, and interest focuses on their temporal association. We propose a joint model for two longitudinal binary processes in which each process is modeled as a nonhomogeneous first-order Markov chain, where the time-dependent transition intensities for each chain depend upon the current and/or past states of the other chain. The joint posterior distribution of the model parameters is obtained using Markov chain Monte Carlo methods. We apply our model to longitudinal data on viral shedding collected from individuals infected with type 2 herpes simplex virus, in which data for each individual were collected from two regions of the genital mucosa.

230 Advances in Bayesian Computation 2 ●

Section on Bayesian Statistical Science, Section on Statistical Computing

Monday, August 1, 2:00 p.m.–3:50 p.m.

Advances In Approximate Bayesian Computation And Trans-Dimensional Sampling Methodology. Savage (Theory And Methods)

✦Gareth William Peters, University of New South Wales, School of Mathematics and Statistics,, UNSW Sydney NSW 2052, Sydney, International 2052 Australia, *garethpeters@unsw.edu.au*

Key Words: approximate Bayesian computation, likelihood-free methodology, equential Monte Carlo, Partial Rejection Control

This talk will present innovations in approximate Bayesian Computation or likelihood-free methodology in two parts. The first section will focus on the technical details for a novel class of Sequential Monte Carlo Samplers algorithms, in which the mutation kernel is modified to incorporate a Partial Rejection Control (PRC) stage. The theoretical and methodological aspects of such a modification are explored. These include detailed specification of the mutation kernel under PRC, including practical and methodological details to efficiently use such a mutation kernel in SMC Samplers methodology. This is particularly relevant to avoiding the calculation of the mutation kernels normalizing constant in the calculation of the incremental importance sampling weights.In addition, specification of the incremental importance sampling weights and analysis of the asymptotic properties of this SMC Samplers PRC algorithm are explored. The second aspect of the talk will discuss developments of multi-variate approximate Bayesian computation methods for some interesting financial and insurance applications.

Abc Methods For Model Choice In Gibbs Random Fields

✦Aude Grelaud, Rutgers University, 509, Hill center, Busch campus, Piscataway, NJ 08854, *agrelaud@stat.rutgers.edu*

Key Words: Model choice, Likelihood free algorithms, Markov random fields

We consider the problem of model selection within the class of Gibbs random fields. In a Bayesian framework, this choice relies on the evaluation of the posterior probabilities of all models. Standard methods of estimation cannot apply in this context since the likelihood is available only up to a normalising constant. We propose here to use a likelihood-free approach to evaluate the posterior model probabilities. We define an extended parameter setting, including the model index.The

ABC-MC algorithm (Approximate Bayesian Computation for Model Choice) generate a sample approximatively distributed from the joint posterior distribution. When all models are Markov random fields, we show the existence of a sufficient statistic for the extended parameter made of the conjunction of the sufficient statistics of all models. This method is then applied to choose the most appropriate 3D structure of a protein among a set of propositions, each candidate corresponding an Ising model with a specific neighborhood structure.

Summary Statistics For Abc

✦Dennis Prangle, Lancaster University, Maths & Stats, Lancaster University, Lancaster, International LA1 4YF UK, *d.prangle@lancaster.ac.uk*; Paul Fearnhead, Lancaster University

Key Words: ABC, Simulation, Queues, Systems biology

Many modern statistical applications involve inference for complex stochastic models, where it is easy to simulate from the models, but difficult or impossible to calculate likelihoods. Approximate Bayesian Computation (ABC) is a method of inference for such models. It replaces calculation of the likelihood by a step which involves simulating artificial data for different parameter values, and comparing summary statistics of the simulated data to summary statistics of the observed data. The results are based on an approximation to the posterior distribution, whose quality depends crucially on the summary statistics. The question of how best to choose these has been an open problem in the literature, with most applications using an ad-hoc choice. This talk describes a generic method to provide summary statistics and presents results showing that it outperforms previous choices in a range of applications including queueing, quantile distributions and systems biology.

231 Methods for Inference from Hard to Rearch Populations ■●

Section on Survey Research Methods, Social Statistics Section, Statistics Without Borders, Committee on Gay and Lesbian Concerns in Statistics

Monday, August 1, 2:00 p.m.–3:50 p.m.

Methods To Infer Hard-To-Reach Populations

✦Richard Sigman, Westat, 1600 Research Blvd, Rockville, MD 20850 USA, *richardsigman@westat.com*; ✦Krista Jennifer Gile, University of Massachusetts, 01003-9305, *gile@math.umass.edu*; ✦Martin Humberto Felix-Medina, UNIVERSIDAD AUTONOMA DE SINALOA, GRAL. A. FLORES PTE S/N, COLONIA CENTRO, CULIACAN SINALOA, International 80000' MEXICO, *mhfelix@uas.uasnet.mx*

Key Words: venue-based sampling, cluster sampling, link-tracing network sampling, hidden population, horvitz-thompson estimator, inference

The study of hard-to-reach or otherwise "hidden" populations presents many challenges to existing survey methodologies. Examples of such populations in a behavioral and social setting include injection drug users, men who have sex with men, and female sex workers. Examples in a broader economic setting include unregulated workers, migrants, homeless, displaced peoples. These populations are characterized by the difficulty in sampling from them using standard probability methods. Typically, a sampling frame for the target population is not available, and its members are rare or stigmatized in the larger population so that it is prohibitively expensive to contact them through the available frames. Hard-to-reach populations in the US and elsewhere are underserved by current sampling methodologies mainly due to the lack of practical alternatives to address these methodological difficulties. This session will review current approaches to the sampling of hidden populations and present advances in statistical methodology.

232 Statistical Methods for the Development of Personalized Medicine

Biometrics Section, ENAR, Section on Health Policy Statistics, WNAR, Scientific and Public Affairs Advisory Committee

Monday, August 1, 2:00 p.m.–3:50 p.m.

A Comparison Of Q- And A- Learning Methods For Estimating Optimal Treatment Regimes

✦Phillip Joel Schulte, North Carolina State University, 601 Hinsdale St. Apt. 3, Raleigh, NC 27605, *pjschult@ncsu.edu*; Marie Davidian, North Carolina State University; Anastasios Tsiatis, North Carolina State University

Key Words: Treatment regimes, Bias-variance tradeoff, Model misspecification, Advantage learning, Q-learning

In clinical practice, physicians must make a sequence of treatment decisions throughout the course of a patient's disease based on evolving patient characteristics. At key decision points, there may be several treatment options and no consensus regarding which option is best. An algorithm for sequential treatment assignment at key decision points, based on evolving patient characteristics, is called a treatment regime. The statistical problem is to estimate the optimal regime which maximizes expected outcome. Q- and A-reinforcement learning are two methods that have been proposed for estimating the optimal treatment regime. While both methods involve developing statistical models for patient outcomes, A-learning is more robust, relaxing some assumptions. However, this additional robustness comes at a cost of increased variability and a bias-variance tradeoff between Q- and A-learning. We explore this tradeoff through parameter estimation and expected outcome for the estimated optimal treatment regime under various scenarios and degrees of model misspecification. We first consider a single treatment decision point for simplicity and then extend to multiple decision points.

Selecting Treatment Strategy Based On Individual Characteristics Using Data From Randomized Clinical Trials

✦Rui Wang, Harvard School of Public Health, 655 Huntington Ave., Building 2, Boston, MA 02115, *rwang@hsph.harvard.edu*; David Alan Schoenfeld, Massachusetts General Hospital

Key Words: clinical trials, treatment effect heterogeneity, qualitative interaction

The primary objective of a Randomized Clinical Trial is to investigate whether one treatment is better than its alternatives on average. However, treatment effects may vary across different patient subpopulations. In contrast to demonstrating one treatment is superior to another on the average sense, one is often more concerned with the question that, for a particular patient, or a group of patients with similar characteristics, which treatment strategy is most appropriate to achieve a desired outcome. We propose a new permutation test for the null hypothesis of no interaction effects for any covariate. The proposed test allows us to consider the interaction effects of many covariates simultaneously without having to group subjects into subsets based on pre-specified criteria and applies generally to randomized clinical trials of multiple treatments. In addition, we introduce a new analytical framework for treatment selection strategy based on individual characteristics. The proposed framework provides one way to generalize the concept of qualitative interaction to multiple covariates in the absence of pre-specified patient subsets.

Treatment Heterogeneity And Individual Qualitative Interaction

✦Robert S. Poulson, Kansas State University, 101 Dickens Hall, Manhattan, KS 66506, *robertp@ksu.edu*; Gary L. Gadbury, Kansas State University; David B. Allison, University of Alabama at Birmingham

Key Words: Causation, Crossover interaction, Individual effects, Potential outcomes, Subject-treatment interaction

Plausibility of high variation in treatment effects across individuals has been recognized as an important consideration in clinical studies. Surprisingly, little attention has been given to evaluating this variability in design of clinical trials or analyses of resulting data. High variation across individuals (referred to herein as treatment heterogeneity) in a treatments efficacy and/or safety across individuals may have important consequences because the optimal treatment choice for an individual may be different from that suggested by a study of average effects. We call this an individual qualitative interaction (IQI), borrowing terminology from earlier work - referring to a qualitative interaction (QI) being present when the optimal treatment varies across "groups" of individuals. We use a potential outcomes framework to elucidate connections between IQI and the probability of a similar response (PSR). We do so under a potential outcomes framework that can add insights to results from usual data analyses and to study design features that improve the capability to more directly assess treatment heterogeneity.

Comparison Of Treatments And Data-Dependent Allocation For Circular Data From A Cataract Surgery

✦Somak Dutta, University of Chicago, Department of Statistics, 5734 S. University Avenue, Chicago, IL 60637, *sdutta@galton.uchicago.edu*; Atanu Biswas, Indian Statistical Institute; Arnab Kumar Laha, Indian Institute of Managemen

Key Words: Circular data, Ethical allocation, Optimal allocation, Response-adaptive design, von Mises distribution

Circular data is a natural outcome in many biomedical studies, e.g. some measurements in ophthalmologic studies, degrees of rotation of hand or waist, etc. With reference to a real data set on astigmatism induced in two types of cataract surgeries we carry out some two-sample testing problems including the Behren-Fisher type of test in the circular set up. Response-adaptive designs are used in phase III clinical trials to allocate a larger proportion of patients to the better treatment. There is no available work on response-adaptive designs for circular data. Here we provide some response-adaptive designs where the responses are of circular nature, first an ad-hoc allocation design, and then an optimal design. Detailed simulation study and the analysis of the data set including redesigning the cataract surgery data are carried out.

Phase Ii Trial Designs To Evaluate Clinical Efficacy Of Personalized Drug Selection Based On An In Vitro Sensitivity Screen

✦Motomi Mori, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, CR145, Portland, OR 97239-3098, *morim@ohsu.edu*; Yiyi Chen, Oregon Health & Science University; Byung S Park, Oregon Health & Science University; Jeffrey W Tyner, Oregon Health & Science University; Marc M Loriaux, Oregon Health & Science University; Tibor Kovacsovics, Oregon Health & Science University; Brian J Druker, Oregon Health & Science University

Key Words: clinical trials, cancer, phase II trials, personalized medicine

We have recently developed a novel in vitro screening technique to evaluate sensitivity of primary cells from leukemia patients against a panel of small-molecule kinase inhibitors. To evaluate the clinical utility of drug selection based on in vitro drug sensitivity, we propose two different randomized phase II trial designs, both involving multiple kinase inhibitors: 1) a single randomization design in which patients are randomized to receive a drug predictive to be effective or one of the other drugs not predictive to be effective, with a higher probability of being assigned to a predicted drug, and 2) a double randomization design in which patients are first randomized to an assay-based drug assignment or randomly selected drug assignment, and those in the randomly selected drug assignment are further randomized to receive one of the drugs under study. We will evaluate operating characteristics of two designs and performance of the proposed frequentist and Bayesian test statistics through simulations. We will conclude our presentation by discussing statistical issues and challenges in testing genome/biomarker-guided strategy or more generally "personalized medicine".

Robust Statistical Method For Finding Optimal Treatment Regimes

✦Baqun Zhang, NCSU, 602 General Joseph Martine, Raleigh, NC 27606, *bzhang4@ncsu.edu*; MARIE DAVIADIAN, NCSU; ANASTASIOS TSIATIS, NCSU

Key Words: Optimal treatment regime, Inverse probability weighting, Potential outcomes, Classification

A treatment regime is a rule that assigns a treatment, among a set of possible treatments, to a patient as a function of his/her observed covariates. The goal is to find the optimal treatment regime defined as that, if followed by a population of patients, would lead to the best outcome on average. For a single treatment decision, the optimal treatment regime can be found by developing a regression model for the expected outcome as a function of treatment and baseline covariates, where, for a given set of covariates, the optimal treatment is the one which yields the largest expected outcome. This, however can lead to biased results

if the regression model is misspecified. Realizing that the parameters in a regression model induce different treatment regimes, we instead consider estimating the mean outcome for such treatment regimes directly using doubly-robust augmented inverse propensity score estimators of the mean outcome for the parameter-induced treatment regimes, which we then maximize across the parameter values to obtain our optimal treatment regime estimator. We also show how this problem can be viewed as a classification problem. Simulations and application are presented

A Policy Search Method For Estimating Treatment Policies

✦Xi Lu, University of Michigan, 439 West Hall, 1085 South University Ave., Ann Arbor, MI 48109, *luxi@umich.edu*; Susan A Murphy, University of Michigan Department of Statistics

Key Words: dynamic treatment regime

A treatment policy or dynamic treatment regime is a sequence of decision rules. At each stage a decision rule inputs patient history and outputs a treatment. The value of a treatment policy is the expected outcome when the policy is used to assign treatment. Data from sequential, multiple assignment, randomized trials can be used to estimate an optimal treatment policy. One approach is to parameterize the policy value (Robins et al. 2008); this may result in bias if the model is misspecified. Alternately the value of any specific policy can be estimated nonparametrically; however this method may have high variance. We propose a new method in which each stage's treatment effect or "blip" is parameterized. These treatment effects are easily interpretable to scientists and thus more meaningfully parameterized than the policy value. To estimate the parameters we utilize a telescoping sum representation of the policy value and uses ideas from missing data theory. We illustrate the proposed method with data from the ExTENd trial, a recently completed alcohol dependence study.

233 Single and Multiple SNP Analysis

Biometrics Section, ENAR

Monday, August 1, 2:00 p.m.–3:50 p.m.

On Model Selection Strategies To Identify Genes Underlying Binary Traits

✦Zheyang Wu, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, *zheyangwu@wpi.edu*; Hongyu Zhao, Yale University

Key Words: Model selection, Genome-wide scan, Statistical power, Marker detection

For more fruitful discoveries of disease genes, it is important to know whether joint analysis of multiple markers is more powerful than the commonly used single-marker analysis, especially in the presence of gene-gene interactions. In studying this problem, the literature has different even conflict arguments about the power of the common model selection strategies: marginal search, exhaustive search, and forward search. In the scenario of binary trait loci detection, we provide a statistical framework to rigorously address the question by analytically calculating the power of these strategies plus a widely used two-stage screen strategy. Our approach incorporates linkage disequilibrium, random genotypes, and correlations among test statistics, which are critical characters of model selection procedures but are often ignored for simplicity in the existing literature. Two types of widely applied error controls were compared: the discovery number control and the Bonferroni type I error rate control. This study provides a fast computation tool in R as well as insights into the statistical mechanism of capturing genetic signals under different epistatic models.

A Bayesian Hierarchical Model For Detecting Haplotype-Haplotype And Haplotype-Environment Interactions In Genetic Association Studies

✦Jun Li, University of Alabama at Birmingham, RPHB 317, 1665 University Boulevard, Birmingham, AL 35294-0022, *junli@uab.edu*; Kui Zhang, University of Alabama at Birmingham; Nengjun Yi, Section on Statistical Genetics, Dept. of Biostatistics, Univ. of Alabama at Birmingham

Key Words: Bayesian methods, Generalized linear models, Genetic association, Hierarchical models, Haplotype-haplotype interactions, Haplotype-environment interactions

Genetic association studies based on haplotypes are powerful in the discovery and characterization of the genetic basis of complex human diseases. However, statistical methods for detecting haplotype-haplotype (HH) and haplotype-environment (HE) interactions have not yet been fully developed. Furthermore, methods for detecting the association between rare haplotypes and disease have not kept pace with their counterpart of common haplotypes. We propose an efficient and robust method to tackle these problems based on a Bayesian hierarchical generalized linear model. Our model simultaneously fits environmental effects, main effects of numerous common and rare haplotypes, and HH and HE interactions. The key to the approach is the use of a continuous prior distribution on coefficients that favors sparseness and facilitates computation. We develop a fast expectation-maximization algorithm to fit models by estimating posterior modes of coefficients. We evaluate the proposed method and compare its performance to existing methods on extensive simulated data. The results show that the proposed method performs well under all situations and is more powerful than existing approaches.

Detecting Natural Selection Across Dependent Populations

✦Eleanne Solorzano, University of New Hampshire, 35 Colovos Road, Durham, NH 03833, *eleannes@cisunix.unh.edu*

Key Words: correlated populations, natural selection, haplotypes

The study of human genes is critical to the understanding of manifestations of diseases. Natural selection is a process that results in the survival and reproductive success of individuals or groups best adjusted to their environment, leading to the perpetuation of genetic qualities best suited to that particular environment. Sabeti's (2002) EHH test has been used to detect recent positive selection of a particular gene across human populations. The standard way to make inferences across the populations is to assume independence using either a Bonferroni or FDR approach. However, human populations are correlated due to the fact that all humans originate from one common African ancestor. Therefore, to reduce bias, it is necessary to account for this correlation among populations. A new statistical method using haplotypes

is developed for detecting natural selection across populations which accounts for such correlations. This test is shown to have higher statistical power than the existing methods to make inferences across populations and also controls the Type I error. The test is illustrated with an example using the lactase gene across 42 populations.

A Gene-Based Supervised Dimension Reduction Approach To Identify Common And Rare Variants In Genome-Wide Association Studies

✦Asuman Turkmen, The Ohio State University, Newark, 1179 University Drive, The Ohio State University at Newark, Newark, OH 43055, *turkmen@stat.osu.edu*; Shili Lin, The Ohio State University

Key Words: genome-wide association studies, rare variants, dimension reduction, PLS, PCA

Genome-wide association studies (GWAS) have become a popular approach for the identification of genes and genetic variants involved in complex diseases. While the standard approach for GWAS has been single-SNP analysis, multi-marker association methods in which multiple markers analyzed jointly to utilize more information have shown to provide more insight into association studies and lead to greater potential in identifying rare associated variants. In this study, we propose to aggregate the signals of many SNPs within a gene using latent components derived by a supervised dimension reduction method by which possible genetic effects related to rare variants can be revealed. Then, a penalized regression model is employed to relate the resulting latent components that represent the genes in the study and the trait of interest to detect associations. The performance of the proposed method is compared with currently available methods.

Rare Or Common? The Clue'S In Cluster.

✦Kaustubh Adhikari, Harvard University, 56 Calumet St., Apt. 2, Boston, MA 02120, *kadhikar.hsph@gmail.com*; Christoph Lange, Harvard University

Key Words: Complex Diseases, Rare Variants, Clustering, Coalescent Trees, Bayesian Modeling

Modern genetic studies involve complex diseases with multiple underlying variants. And it is of interest to know if there are a few common variants or many rare variants. This paper presents a new, clustering-based method where, based on SNP data on cases and controls, we construct random clusters of the subjects - we utilize a novel discovery-based cluster ensemble method by Grimmer and King to sample through the cluster space. The clustering probabilities assist us to construct coalescent genealogical trees for all the subjects, which are then aggregated in a Bayesian framework to provide a posterior distribution on the variants. We apply this method on a cleft lip dataset to show that the posterior distribution indicates that the candidate gene is enriched in rare variants.

A New Robust Method For Testing Single Snp Association In Gwa Studies

✦Zhongxue Chen, The University of Texas Health Science Center at Houston, 77030 USA, *Zhongxue.Chen@uth.tmc.edu*; Tony Ng, Southern Methodist University

Key Words: SNP, GWA, trend test, Chi-square test

In genetic association studies, due to the varying underlying genetic models, there exists no single statistical test that is most powerful under all situations. Current studies show that if the underlying genetic models are known, trend-based tests, which outperform the classical Pearson's chi-square test, can be constructed. However, when the underlying genetic models are unknown, chi-square test is usually more robust than trend-based tests. In this paper, we propose a new statistical testing procedure which is based directly on the disease relative risks, which are order-restricted. Through a Monte Carlo simulation study, we show that this new method is generally more powerful than the chi-square test, and more robust than trend test. The proposed methodologies are illustrated by some real datasets.

Nonparametric Failure Time Analysis With Genomic Applications

✦Cheng Cheng, St. Jude Children's Research Hospital, Department of Biostatisitcs, 262 Danny thomas Place, Memphis, TN 38105, *cheng.cheng@stjude.org*

Key Words: Survival Analysis, rank correlation, GWAS, proportional hazard, permutation test, nonparametric

Genome-wide Association Study (GWAS) has become routine in cancer genomic translational research, which often requires genome-wide screening to identify ordinal genomic features that are associated with treatment outcome, for example, single nucleotide polymorphisms associated with time to relapse. The estimated coefficient of a hazard rate regression model (HRRM) is often used as the association test statistic. It will be demonstrated in this talk that in certain cases the HRRM approach is problematic. A robust, completely nonparametric alternative using rank correlation is then proposed. This method, called correlation profile test (CPT), consists of the correlation profile statistic and a very efficient hybrid permutation test combining permutation and asymptotic theory. Statistical performances are compared with several established methods, by a simulation study and analysis of real genomics data. It is shown that CPT performs much better than the HRRM approach in terms of maintaining the power and nominal significance level, especially in cases where the proportional hazard model does not hold.

234 Bayesian Modeling in Life Sciences 3

Section on Bayesian Statistical Science

Monday, August 1, 2:00 p.m.–3:50 p.m.

A Bayesian Nonparametric Method For Differential Expression Analysis Of Rna-Seq Data

✦Yiyi Wang, Department of Statistics, Texas A&M University, College Station, TX 77840, *wangyiyi@stat.tamu.edu*; David B. Dahl, Department of Statistics, Texas A&M University

Key Words: RNA-seq, Gene Ontology, Differential expression, Bayesian nonparametric, Next generation sequencing

We developed a method for RNA-seq data to identifying differentially expressed genes. Our method builds upon the negative binomial model of others, but uses Gene Ontology (GO) annotations as prior information to assist the clustering process and to improve sensitivity and specificity. Our method allows each gene to have its own parameters and estimates those parameters by a Bayesian nonparametric technique which shares information across genes in the same cluster. The method explicitly calculates the probability that each gene is differentially expressed and the genes are ranked by these probabilities. For any set of genes having high probability of differential expression, the estimated false discovery rate is computed. Thresholds can be adjusted to achieve a desired estimated false discovery rate. We demonstrated with an actual data set.

Integrative Analysis Of Dna Copy Number And Gene Expression Data

✦Runqi Lin, Southern Methodist University, 5349 Amesbury Drive Apt#510, Dallas, TX 75206 USA, *rlin@smu.edu*; Xinlei (Sherry) Wang, Southern Methodist University; Guanghua (Andy) Xiao, UT Southwestern Medical Center

Key Words: Bayesian methods, Comparative genomic hybridization (CGH), Single nucleotide polymorphism (SNP), Hidden Markov model (HMM)

Array comparative genomic hybridization (CGH) and single nucleotide polymorphism (SNP) arrays can be used to detect genomic DNA copy number alterations, which are closely related to the development and progression of cancer. Such alterations, including amplifications and deletions, can result in significant changes in gene expression. A gene is called a tumor driver gene if its copy number variation leads to the change in gene expression, and hence plays a key role in tumor genesis. Integrative analysis of the copy number data (i.e., the array CGH or SNP data) and gene expression data simultaneously can not only improve the statistical power for identifying the tumor driver genes, but also provide a comprehensive picture of biological mechanisms. While a large number of approaches have been proposed to analyze the copy number data alone, there is still lack of statistical methodology for integrative analysis of the copy number and gene expression data. In this paper, we will adopt a Bayesian approach relying on the hidden Markov model (HMM) to incorporate the information from the gene expression and copy number data and identify tumor driver genes.

A Fully Bayesian Hidden Ising Model For Chip-Seq Data Analysis

✦Qianxing Mo, Memorial Sloan-Kettering Cancer Center, 307 E 63rd Street, 3rd Floor, New York, NY 10065, *moq@mskcc.org*

Key Words: ChIP-seq, Ising model, next generation sequencing, Massively parallel sequencing, Bayesian Hierarchical, Transcription factor

Chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) is a powerful technique that is being used in a wide range of biological studies. To systematically model ChIP-seq data, we build a dynamic signal profile for each chromosome, and then model the profile using a fully Bayesian hidden Ising model. The proposed model naturally takes into account spatial dependency, global and local distributions of sequence tags. It can be used for one-sample and two-sample analyses. Through model diagnosis, the proposed method can detect falsely enriched regions caused by sequencing and/or mapping errors, which is usually not offered by the existing hypothesis-testing-based methods. The proposed method is illustrated using three transcription factor ChIP-seq data sets and four mixed ChIP-seq data sets, and compared with four popular and/or well-documented methods: MACS, CisGenome, BayesPeak and SISSRs. The results indicate that the proposed method achieves equivalent or higher sensitivity and spatial resolution in detecting transcription factor binding sites with false discovery rate at a much lower level.

Bayesian Variable Selection For High-Dimensional Clustering

✦Juanjuan Li, Northwestern University, 2006 Sheridan Road., Evanston, IL 60208, *juanjuanli2008@u.northwestern.edu*

Key Words: Bayesian variable selection, Clustering analysis, Gibbs sampler

Variable selection and parameter reduction are two challenges for high-dimensional clustering with finite mixture models. Many of the previous variable selection procedures assume that the non-discriminating variables are independent of the discriminating variables. A more general variable selection model is proposed by assuming that some of the non-discriminating variables are linearly associated with the discriminating variables. A Bayesian algorithm is derived for this new model which can simultaneously conduct variable selection and clustering. For parameter reduction, the mean, covariance, and regression parameters of the model are integrated out and the marginal posterior function is used for the MCMC algorithm for the rest of the clustering parameters. The posterior inference and model identification are specified. Also the performance of the methodology is explored with simulated and real datasets.

Bayesian Modeling For Histone Modifications

✦Ritendranath Mitra, MD Anderson Cancer Centre, 7901 Cambridge street apt 53, Houston, TX 77054 United States, *riten82@gmail.com*

Key Words: Auto-logistic, MCMC, Pathway dependence, Markov random fields

Histone modifications (HMs) are an important post-translational feature. Different types of HMs are believed to co-regulate biological processes such as gene expression, and therefore are intrinsically dependent on each other. We develop inference for this complex biological network of HMs based on a graphical model for the dependence structure across HMs. A critical computational hurdle in the inference for the proposed graphical model is the evaluation of a normalization constant in an autologistic model that builds on the graphical model. We tackle the problem by constructing posterior Markov chain Monte Carlo (MCMC) simulation in a way that avoids the evaluation of these normalization constants. We report inference on HM dependence in a case study with ChIP-Seq data from a next-generation sequencing experiment. An important feature of our approach is that we can report coherent probabilities and estimates related to any event or parameter of interest, including honest uncertainties. Posterior inference is obtained from a joint probability

Meta-Analysis Of Functional Neuroimaging Data Using Bayesian Nonparametric Binary Regression

✦Yu Yue, Baruch College, The City University of New York, One Bernard Baruch Way, New York, NY 10010, *yu.yue@baruch.cuny.edu*; Martin A Lindquist, Columbia University; Ji Meng Loh, AT&T Labs-Research

In this work we perform a meta-analysis of neuroimaging data, consisting of locations of peak activations identified in 162 separate studies on emotion. Neuroimaging meta-analyses are typically performed using kernel-based methods. However, these methods require the width of the kernel to be set a priori and to be constant across the brain. To address these issues, we propose a fully Bayesian nonparametric binary regression method to perform neuroimaging meta-analyses. In our method, each location (or voxel) has a probability of being truly activated, and the corresponding probability function is based on a spatially adaptive Gaussian Markov random field (GMRF). We also include parameters in the model to robustify the procedure against miscoding of the voxel response. Posterior inference is implemented using efficient MCMC algorithms extended from those introduced in Holmes and Held (2006). Our method allows the probability function to be locally adaptive with respect to the covariates, that is, to be smooth in one region of the covariate space and wiggly or even discontinuous in another. Posterior miscoding probabilities for each of the identified voxels can also be obtaine

235 Advances in Process Control ●

Section on Quality and Productivity, Section on Physical and Engineering Sciences

Monday, August 1, 2:00 p.m.–3:50 p.m.

Sampling Strategies to Reduce the Effect of the Autocorrelation on the Xbar Chart's Performance

✦Marcela Machado, UNESP, Guaratinguet·, 12516-410 Brazil, *marcela@feg.unesp.br*; Antonio Costa, UNESP

Key Words: autocorrelation, control chart, sampling strategies

In many fast production lines, the neighboring units, regarding to the time they were produced, are highly correlated, that is, the X values of the quality characteristic under surveillance from neighboring items are autocorrelated. According to the literature, the dependence among neighboring items fits pretty well to an autoregressive model AR (1). We adopt this model in our study and we build-up the samples with neighboring items considering the rational subgroup concept. The consequence of the autocorrelation among sample units is a drastic reduction in the Xbar chart's ability in signaling process mean changes. The first sampling strategy to reduce the effect of the autocorrelation consists in building-up the sample by taking one item from the production line and skipping one, two or more before selecting the next. The second strategy consists in working with mixed samples, that is, we build-up the mixed sample with units from the current sample plus items from the former or the formers samples. The main conclusion we take from this study is that the proposed sampling strategies counteract the undesired effect of the autocorrelation.

Evaluation Of The Performance Of A Random Coefficient Regression Model Cusum Control Chart Under Varying Model Conditions: With Human Services Applications

✦Christopher John McKinney, University of Northern Colorado, 1554 Peacock Place, Loveland, CO 80537 USA, *cj.mckinney@live.com*; Jay Schaffer, University of Northern Colorado

Key Words: CUSUM, random coefficient regression model, regression control chart, human services

The use of quality control charts with metrics within the educational, behavioral, and other human services field has typically been considered very limited due to the complications imposed by nested structures, conditional relationships, and complex variance-covariance structures. The current study demonstrates how under these population conditions the random coefficient regression model control chart (RCRMCC) is a useful alternative traditional control charts. Derived from the Regression Control Chart (RCC), the RCRMCC utilizes the residuals of the random coefficient regression model to provide the inputs for the quality control charts. Under the CUSUM framework and using Monte Carlo simulations, the RCRMCC and RCC are compared in regards to their ARLS under varying in-control and out-of-control population conditions. It was found that the RCRMCC is advantageous under conditions similar to those found in the educational, behavioral, and other human services industries. Applications of the findings and future directions with the RCRMCC are further discussed.

Monitoring Variability Of Multivariate Processes

✦Amit Mitra, Auburn University, College of Business, Suite 516, Auburn, AL 36849-5240, *mitraam@auburn.edu*

Key Words: Multivariate process control, Variance-covariance matrix, time to first detection

Techniques to monitor and detect shifts in the mean of a multivariate process have received much attention. Here, the focus is on detecting and possibly identifying changes in the variability of response variables in multivariate processes. Traditional methods incorporate the determinant of the variance-covariance matrix of the response variables. This has an advantage of aggregating the variability of several variables into one index. However, it also has a drawback of failing to detect changes in variances since it is quite possible for different for different variance-covariance matrices to yield the same determinant. Decomposition of the variance-covariance matrix is investigated that may lead to better identification of changes in the variances of the response variables. Results of the suggested procedure are investigated through simulation. The mean and standard deviation of the time to first detection, in the event of a change, are used as performance measures for detection out-of-control conditions.

Phase I Control Chart Based On A Kernel Estimator Of The Quantile Function

✦Gary R Mercado, The University of Alabama, 1108 14th Ave Apt 325, Tuscaloosa, AL 35401, *grmercadovelasco@crimson.ua.edu*; Michael D. Conerly, The University of Alabama; Marcus Perry, The University of Alabama

Key Words: kernel estimation, bandwidth, root mean square error, average run length, Shewhart control chart

To measure the statistical performance of a control chart in Phase I applications, the in-control average run length (ARL) is the most frequently used parameter. For start up situations, control limits must be computed without knowledge of the underlying distribution of the quality characteristic, in many cases. Assumptions of an underlying normal distribution are not always fulfilled and can increase the probability of false alarms which can lead to unnecessary process adjustments. In this talk, a control chart based on a kernel estimator of the quantile function is suggested. Monte Carlo simulation and the mean absolute index (MAI) were used to evaluate the in-control ARL performance of this chart relative to that of the Shewhart individuals control chart. Results indicate that the proposed chart is robust with respect to the in-control ARL and results in an alternative method of designing control charts for individual units.

Stability Analysis In The Exponential Families And Generalized Linear Model

✦Ying Lu, University of Minnesota, 224 Church St SE, 313 Ford Hall, Minneapolis, MN 55455, *luxxx255@umn.edu*

Key Words: Cumulative Sum, Sequential Change Detection, Stability, Exponential Family, Generalized Linear Model

In many modern applications of Statistics,the data arrives in a sequential order, and the question of interest is whether the parameters of the data generating system have remained stable through the entire time-length of data collection. We construct a likelihood ratio test statistic to study the stability of parameters, which turns out to be the well-known CUSUM statistic. We study the effectiveness of using our distribution-specific test statistic, as opposed to the normality-driven CUSUM statistic. The conclusion is that when the underlying distribution belongs to the exponential family, our likelihood-ratio statistic generally works better than the normal CUSUM statistic, especially when the potential change in parameters is small in size. Another observation is that the likelihood-ratio statistic has a stable performance in detection whether the change point lies at the beginning, in the middle or at the end of the data collection time. We extend our study to test for changes in parameters in the generalized linear model setting. Finally, we conduct a detailed study on the Atlantic hurricane data over the past 150 years, and the validity of our methodology is strengthened.

236 Ensemble Methods ●

Section on Statistical Computing, Section on Statistical Graphics
Monday, August 1, 2:00 p.m.–3:50 p.m.

Pruning Ensemble Models For Classification Problems On Large Datasets

✦Damir Spisic, IBM USA, 233 S. Wacker Drive, Floor 11, Chicago, IL 60606, *dspisic@us.ibm.com*; Fan Li, IBM China; Jing Xu, IBM China

Key Words: classification, ensemble models, ensemble pruning, large data modeling

Classification methods that generate a single model are generally not suitable for analysis of very large datasets, streaming data or updated data. We consider an ensemble modeling approach where component models are generated on disjoint consecutive data blocks. This approach is efficient and effectively addresses all three stated challenges. Several learning methods are used for creating component models and a few different combination methods are considered for computing ensemble predictions. Ensembles are pruned in order to keep the number of component models limited and therefore efficient for model scoring. We study and compare several pruning methods in this context using validation and test samples extracted from the overall dataset. Both accuracy and diversity measures are considered in this context. We assume stationary and randomly ordered datasets and show that this approach is not only efficient and scalable, but also produces similar or superior prediction accuracy when compared to the single model generated by the same learning method.

A Quadratic Programming Algorithm For Reducing The Size And Improving The Performance Of An Ensemble Model

✦Jie Xu, University of Alabama, ISM Dept, 300 Alston Hall, Box 870226, Tuscaloosa, AL 35487-0226, *xjpat2046@gmail.com*; J. Brian Gray, University of Alabama

Key Words: bagging, boosting, machine learning, predictive modeling, random forests

Ensemble models, such as bagging, random forests, and boosting, have better predictive accuracy than single classifiers. These ensembles typically consist of hundreds of single classifiers, which makes future predictions and model interpretation much more difficult than for single classifiers. According to Breiman (2001), the performance of an ensemble model depends on the strengths of the individual classifiers in the ensemble and the correlations among them. In this article, we propose a new method based on quadratic programming that uses information on the strengths of, and the correlations among, the individual classifiers in the ensemble, to improve or maintain the predictive accuracy of an ensemble while significantly reducing its size.

The Interactive Decision Committee For Chemical Toxicity Analysis

✦Chaeryon Kang, Dept. of Biostatistics, The University of North Carolina at Chapel Hill, 109 Timber Hollow court APT #344, Chapel Hill, NC 27514, *ckang@bios.unc.edu*; Hao Zhu, The University of North Carolina at Chapel Hill; Fred Andrew Wright, Univ North Carolina; Fei Zou, The University of North Carolina at Chapel Hill ; Michael R Kosorok, UNC-CH

Key Words: Chemical toxicity, Decision committee method, Ensemble, Ensemble feature selection, QSAR modeling, Statistical learning

We introduce the Interactive Decision Committee method for classification when high-dimensional feature variables are grouped into feature categories. The proposed method uses the interactive relationships among feature categories to build base classifiers which are combined using decision committees. A two-stage 5-fold cross-validation technique is utilized to decide the total number of base classifiers to be combined. The proposed procedure is useful for classifying biochemicals on the basis of toxicity activity, where the feature space consists of chemical descriptors and the responses are binary indicators of toxicity activity. Each descriptor belongs to at least one descriptor category. The support vector machine algorithm is utilized as a classifier inducer. Forward selection is used to select the best combinations of the base classifiers given the number of base classifiers. We applied the pro-

posed method to two chemical toxicity data sets. For these data sets, the proposed method outperforms other decision committee methods including adaboost, bagging, random forests, the univariate decision committee, and a single large, unaggregated classifier.

On The Insufficiency Of The Large Margins Theory In Explaining Boosting And Ensemble Performance

✦Waldyn Martinez Cid, University of Alabama, ISM Dept, 300 Alston Hall, Box 870226, Tuscaloosa, AL 35487-0226, *wmartine@cba.ua.edu*; J. Brian Gray, University of Alabama

Key Words: AdaBoost, arc-gv, generalization error, linear programming

AdaBoost (Freund and Schapire 1997) and other ensemble methods combine a set of weak classifiers through weighted voting to produce a strong classifier. Schapire, Freund, Bartlett, and Lee (1998) developed a bound on the generalization error of a combined classifier based on the margins of the training data, the sample size, and the complexity of the weak classifiers. From this bound, they and others (see, e.g., Reyzin and Schapire 2006) have concluded that higher margins should lead to lower generalization error, everything else being equal (sometimes referred to as the "large margins theory"). In this article, we introduce a linear programming (LP) method that increases or maintains all of the margins of an AdaBoost (or any other ensemble) solution using the same set of weak classifiers, yet the resulting combined classifier has similar or worse test set performance than AdaBoost, indicating that the large margins theory is insufficient to explain the performance of AdaBoost.

Determining Fitness Function Parameters For Ga-Boost

✦Dong-Yop Oh, University of Alabama, ISM Dept, 300 Alston Hall, Box 870226, Tuscaloosa, AL 35487-0226, *doh@cba.ua.edu*; J. Brian Gray, University of Alabama

Key Words: AdaBoost, classification, genetic algorithm, predictive model, weak classifier

Our recently proposed genetic boosting algorithm, GA-Boost, directly solves for the weak classifiers in an ensemble and their weights using a genetic algorithm. The fitness function consists of three parameters (a, b, and p) that limit the number of weak classifiers (by b) and control the effects of outliers (by a) to maximize an appropriately chosen p-th percentile of margins. We use several artificial data sets to compare GA-Boost performance at 16 different treatment levels, as well as how it compares to AdaBoost, at four different noise levels. Through these simulations, we verify that GA-Boost has better performance with simpler predictive models than AdaBoost when there is a large proportion of outliers in a data set. GA-Boost is applied to real data sets with three different weak classifier options and compared to other robust boosting methods. We also consider graphical methods for selecting the value of p.

Strategies For Extracting Knowledge From Ensemble Classifiers Based On Generalized Additive Models

✦Koen W. De Bock, IESEG School of Management (Lille & Paris, France), 3, Rue de la Digue, Lille, International 59000 France, *k.debock@ieseg.fr*

Key Words: ensemble classification, generalized additive models, GAMens, model interpretability, bagging

In recent literature, ensemble learning has demonstrated superior performance in a multitude of applications. However, their increased complexity often prevents qualitative model interpretation. In this study, GAMensPlus, an ensemble classifier based upon generalized additive models (GAMs), in which both performance and interpretability are reconciled, is presented and evaluated. The recently proposed GAMens, based upon Bagging, the Random Subspace Method and semi-parametric GAMs as constituent classifiers, is extended to include two instruments for model interpretability: generalized feature importance scores, and bootstrap confidence bands for smoothing splines. These elements allow insight into (i) the relative importance of features within the model, (ii) the nature of the relationship between each feature and the outcome, and (iii) the reliability of these estimated trends at different regions within the range of the feature. In an experimental comparison on UCI and simulated datasets, the strong classification performance of the proposed algorithm versus a set of well-known benchmark algorithms as well as properties of the strategies for interpretability are demonstrated.

Gibbs Ensembles For Nearly Compatible And Incompatible Conditional Models

✦Shyh-Huei Chen, Wake Forest University School of Medicine, NC 27157, *schen@wfubmc.edu*; Edward H Ip, Wake Forest University School of Medicine; Yuchung J Wang, Rutgers University

Key Words: Gibbs sampler, Conditionally specified distribution, Linear programming, Ensemble method; Odds ratio

The Gibbs sampler has been used exclusively for compatible conditionals that converge to a unique invariant joint distribution. However, conditional models are not always compatible. In this paper, a Gibbs sampling-based approach-using the Gibbs ensemble-is proposed for searching for a joint distribution that deviates least from a prescribed set of conditional distributions. The algorithm can be easily scalable, such that it can handle large data sets of high dimensionality. Using simulated data, we show that the proposed approach provides joint distributions that are less discrepant from the incompatible conditionals than those obtained by other methods discussed in the literature. The ensemble approach is also applied to a data set relating to geno-polymorphism and response to chemotherapy for patients with metastatic colorectal cancer.

237 Survival Analysis

Section on Nonparametric Statistics, ENAR, International Indian Statistical Association

Monday, August 1, 2:00 p.m.–3:50 p.m.

Novel Maximum Spacing for Point Estimates from a Semiparametric Ratio Estimator: With Application to Survival in Prostate Cancer

✦Deborah Weissman-Miller, Brenau University, 500 Washington Street, SE, Gainesville, GA 30501, *dweissman-miller@brenau.edu*

Key Words: Spacing estimator, semiparametric ratio estimator, change point, clinical trial, prostate cancer

Prostate cancer is a condition of public health significance in the United States. A novel maximum spacing derivation is introduced to determine the maximum time intervals between point estimates for a new semiparametric ratio estimator to predict survival in response to treatment for prostate cancer. A maximum spacing estimation method is derived from a univariate distribution where spacing will be defined as gaps between ordered values of the distribution function. The spacing estimator uses the geometric mean of sample spacings from a uniform distribution U(a,b) with one known endpoint given as a change point derived from the highest or lowest F statistic from a linear regression for the semiparametric ratio estimator. The maximum is defined as a single value in the neighborhood of the change point and the spacing defined as a function of time. This maximum spacing then defines the gaps between point estimates at each time-dependent predicted outcome from the change point and results in a semiparametric ratio estimator that is reliable and repeatable. A real application of this maximum spacing estimator is to predict the survival of prostate cancer patients in clinical medicine.

A New Nonparametric Estimator Of A Survival Function

✦Ganesh B Malla, Xavier University, 8004 Higgins Ct, Cincinnati, OH 45242 USA, *mallag@xavier.edu*; Hari G. Mukerjee, Wichita State University

Key Words: Piecewise exponential estimator, Kaplan-Meier estimator, Survival function, Hazard rate

In 1983, Kitchin, Langberg and Proschan introduced a piecewise exponential estimator (PEXE) of a survival function (SF) for censored data. It provides a continuous estimator of the SF as well as a new method of handling the censoring. This PEXE has had limited usage possibly because the estimator is undened beyond the last bservation even in the uncensored case. We propose a new PEXE that retains the spirit of the Kaplan-Meier estimator and provides an exponential tail with a hazard rate determined by a novel nonparametric consideration. A comparison, by simulation, of the model with other models has been considered.

Crossing Hazard Functions In Common Survival Models

✦Jiajia Zhang, University of South Carolina, Department of Epidemiology and Biostatistics, Arnold School of Public Health, 800 Sumter Street, Columbia, SC 29201, *jzhang@mailbox.sc.edu*; Yingwei Peng, Queen's University

Key Words: Accelerated hazard model, Accelerated failure time model, monotone hazard, U-shape hazard, Bell-shape hazard

Crossing hazard functions have extensive applications in modeling survival data. However, existing studies in the literature mainly focus on comparing crossed hazard functions and estimating the time at which the hazard functions cross, and there is little theoretical work on conditions under which hazard functions from a model will have a crossing. In this paper, we investigate crossing status of hazard functions from the proportional hazards (PH) model, the accelerated hazard (AH) model, and the accelerated failure time (AFT) model. We provide and prove conditions under which the hazard functions from the AH and the AFT models have no crossings or a single crossing. A few examples are also provided to demonstrate how the conditions can be used to determine the crossing status of hazard functions from the three models.

Negative Moment Inequalities Of Life Distributions With Hypothesis Testing Applications: Drhr, Drhrs, Imit, And Imits Classes

✦Mohammad B Sepehrifar, The University of Mississippi, P. O. Box 848, University, MS 38677, *moe@olemiss.edu*

Key Words: U-statistic, Negative moment inequality, Reversed hazard rate, Inactivity time, Exponentiality

Ibrahim and Sepehrifar [1] studied both probabilistic and statistical properties of some new aging classes of life distributions including DRHR, and IMITS. The purpose of this talk is to introduce the new negative moment inequalities for the aforementioned aging classes of life distributions. These inequalities enable us to perform some nonparametric procedures for testing family of gamma against any alternative distributions belong to one of the above classes. These tests are based on sample negative moments of these aging classes; they are simple to devise, to calculate and to study relative to other more complicated tests in the literature. The limiting distributions of the presented test statistics are given for a well known alternative when the null distribution belongs to a family of gamma. The Monte Carlo study shows an excellent power of these procedures for some common alternative distributions.

Cure Rate Models With Partially Observed Covariates

✦Tzu-Chun Lin, University of California, Davis, , *tclin@wald.ucdavis.edu*

Key Words: cure rate model, partially observed covariates, missing value, nonparametric baseline hazard

Cure rate models are proposed to model time-to-event data with a significant fraction of subjects who are considered as long-term survivors. One design of these models is a mixture of a cure probability and a survivor function for the non-cured group with a non-cured probability. Meanwhile, partially observed covariates data are often encountered due to many reasons, and analysis only based on complete data will often result in bias estimates. This study aims to compare a nonparametric baseline hazard logistic/PH mixture cure rate model with a commonly chosen model, a piecewise constant baseline hazard logistic/PH mixture cure rate model with covariates having missing values.

A Nonparametric Test For Equality Of Survival Medians

✦Mohammad Hossein Rahbar, University of Texas Health Science Center at Houston, UTPB Room 11.05, 6410 Fannin Street, Houston, TX 77030, *Mohammad.H.Rahbar@UTH.TMC.EDU*; Zhongxue Chen, The University of Texas Health Science Center at Houston; Sangchoon Jeon, Yale School of Nursing; Joseph C Gardiner, Michigan State University; Jing Ning, The University of Texas School of Public Health

Key Words: Median, Nonparametric, Censored Data, Clinical Trials, Survival Analysis, Simulations

In clinical trials researchers often encounter testing for equality of survival medians across study arms based on censored data. Even though Brookmeyer-Crowley (BC) introduced a method for comparing medians of several survival distributions, still some researchers misuse procedures which are designed for testing the homogeneity of survival curves. This practice often leads to inflation of type I error. We propose a new nonparametric method for testing the equality of several survival medians based on Kaplan-Meier estimation from randomly right censored data. We derive asymptotic properties of this test statistic. Through simulations we compare the power of this new procedure with that of the log-rank, Wilcoxon, and BC method. Our simulation results indicate that performance of these test procedures depends on the level of censoring and appropriateness of the underlying assumptions. When the assumptions of the log-rank test are met, some of these procedures are more powerful than the Wilcoxon, BC, and our proposed test. However, when the objective is testing homogeneity of survival medians rather than survival curves, our test statistic provides an alternative to the BC test.

238 Computer Experiments ■●

Section on Physical and Engineering Sciences, International Indian Statistical Association, Section on Quality and Productivity

Monday, August 1, 2:00 p.m.–3:50 p.m.

Designs For Ensemble Of Multiple Computer Models

✦Youngdeok Hwang, University of Wisconsin-Madison, Department of Statistics, 1300 University Ave, Madison, WI 53706, *hwang@stat.wisc.edu*; Peter Z. G. Qian, University of Wisconsin-Madison

Key Words: Computer Experiments, U designs, Space-Filling Designs, Sliced Designs

Ensemble of multiple computer models arises as an important tool to study the complex phenomena and is getting more traction in many disciplines. We propose a new type of design, called sliced U design, for running such models. A sliced U design is a U design based on an orthogonal array with strength two that can be divided into different slices, each of which is a Latin hypercube design associated with each individual model. The construction method is easy to implement. Also will be discussed is a two-step procedure for integrating data from multiple computer models for estimating the average output mean. Numerical illustration is accompanied to validate the derived theoretical results.

A Composite Likelihood Approach For Computer Experiments

✦Ximing Xu, University of Toronto, Toronto, ON m5t 1m7 Canada, *siemenxu@gmail.com*; Derek Bingham, Simon Fraser University; Nancy Reid, University of Toronto

Key Words: computer experiment, composite lilkelihood, universal kriging, Gaussian process

Modeling the computer outputs as a sample path of a Gaussian process has proven useful to predict the outputs at untried input sites. However as the number of simulator evaluations gets larger, statistically modeling the computer outputs usually becomes computationally intractable. In this paper we model the mean of the Gaussian process as a regression function of the input vector and estimate the unknown parameters based on the pairwise likelihood of the high dimensional Gaussian distribution. We also derive a new method to do prediction by maximizing a special composite likelihood. The composite likelihood approach does not need to calculate the inverse or determinant of the high dimensional covariance matrix and require fewer model assumptions and much less computation time than full likelihood based methods. Finally our approach is applied to analyze a real photometric redshift data and performs very well in predictive accuracy.

Modeling Mesh Density In Computer Experiments

✦Rui Tuo, School of Industrial and Systems Engineering, Georgia Institute of Technology. AMSS, CAS., 765 Ferst Drive, NW, Atlanta, GA 30332-0205, *tuorui@amss.ac.cn*; C. F. Jeff Wu, School of Industrial and Systems Engineering, Georgia Institute of Technology; Dan Yu, Academy of Mathematics and SystemsScience, Chinese Academy of Sciences

Key Words: kriging, multi-resolution data, nonstationary guassian process models, tuning parameters

In a computer experiment a set of partial differential equations is solved numerically to simulate the result of a corresponding physical experiment. The accuracy of a numerical algorithm such as the finite element method is determined by the mesh density. As the mesh density increases, the numerical accuracy is improved, while the computational cost goes up. The quantitative relationship between error and mesh density is available in the theory of finite element. New nonstationary Gaussian process models are proposed to establish a framework consistent with the results in numerical analysis. These models can be analyzed by Bayesian methods. The proposed method is also applicable to multi-resolution experiments. The methodology is illustrated by two examples.

Multi-Layer Designs For Computer Experiments

✦Shan Ba, School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive, NW, Atlanta, GA 30332, *shan.ba@gatech.edu*; Roshan Joseph Vengazhiyil, School of Industrial and Systems Engineering, Georgia Institute of Technology

Key Words: Foldover technique, Fractional factorial design, Latin hypercube design, Minimum aberration criterion, Space-filling design

Space-filling designs such as Latin hypercube designs (LHDs) are widely used in computer experiments. However, finding an optimal LHD with good space-filling properties is computationally cumbersome. On the other hand, the well-established factorial designs in physical experiments are unsuitable for computer experiments owing to the redundancy of design points when projected onto a subset of factor space. In this work, we present a new class of space-filling designs developed by splitting two-level factorial designs into multiple layers. The method takes advantages of many available results in factorial design theory and therefore, the proposed Multi-layer designs (MLDs) are easy to generate. Moreover, our numerical study shows that MLDs can have better space-filling properties than optimal LHDs.

Bayesian Emulators For Multivariate Computer Models With Categorical Inputs

✦David Woods, Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, Highfield, Southampton, International SO17 2BJ United Kingdom, *davew@soton.ac.uk*; Antony Overstall, Southampton Statistical Sciences Research Institute (S3RI)

Key Words: Computer experiment, Gaussian Process, Bayesian statistics, Markov Chain Monte Carlo Model Composition, Sensitivity analysis, Dispersion modelling

Computer models, or simulators, are mathematical representations of physical systems. They are used for real-world problems where it would be expensive, impossible or unethical to use a physical experiment. Many simulators are computationally expensive and hence an emulator, a statistical meta-model, is used to predict the simulator output at any particular set of inputs. Emulators make practicable tasks such as sensitivity analysis, uncertainty analysis and model calibration. We consider simulators for real problems from emergency planning that have several defining features: the output is multivariate, potentially dynamic and may be zero-inflated; the input may include both continuous and categorical input variables. Challenges for such models include the definition and incorporation of appropriate distance metrics for the categorical variables and implementing efficient methods for approximating the multivariate posterior predictive distribution. We explore a variety of Bayesian methods for constructing emulators, including Gaussian Process regression and the use of "lightweight" emulators, and address issues including predictive accuracy and model selection.

239 Missing Data Analysis ■

Biopharmaceutical Section

Monday, August 1, 2:00 p.m.–3:50 p.m.

Comparing Mixed Models, Pattern Mixture Models, And Selection Models In Estimating Treatment Effects Of Quality Of Life Data With Nonignorable Missing Data In Clinical Trial Study

✦Xiaolei Zhou, RTI Health Solutions, Research Triangle Park, NC 27709, *xzhou@rti.org*; Jianmin Wang, RTI Health Solutions; Jessica Zhang, RTI Health Solutions; Hongtu Zhu, University of North Carolina Department of Biostatistics

Key Words: missing data, longitudial data analysis, simulation, bias

Missing data commonly occur in health-related quality of life data in clinical trial studies because these data are reported by patients. Mixed effects models are frequently chosen for the primary analysis of clinical trial studies. However, mixed effects models require data missing at random. Pattern mixture models and selection models are two methods for handling nonignorable missing data. Each method is built on specific assumptions on the mechanism of missing data. We are interested in the magnitude of the bias and the robustness of mixed effects models, pattern mixture models, and selection models under various missing mechanisms. Simulations were performed that focused on the treatment effect, which is the primary interest of clinical trial studies. Simulation data were generated from 10 missing mechanisms that may occur in clinical trial studies. Both the point estimate and the variance of the estimate were calculated for each model. These analyses are the first we know to systematically evaluate and compare the mixed effects models, pattern mixture models, and selection models using simulation data.

An Application Of Multiple Imputation In Analyses Of Longitudinal Clinical Trials With Left Truncated Data

✦Guanghan Liu, Merck Research Laboratories, 351 N. Sumneytown Pike, North Wales,, PA 19454 United States, *guanghan_frank_liu@merck.com*

Key Words: Multiple imputation, Longitudinal analysis, Left truncation

For analysis of lab measurements collected in clinical trials, the outcomes may suffer from left truncation at the detection limit of the lab assay. In the conventional methods, the left-truncated values are often replaced with the detection limit or half of the detection limit. This simple single imputation could be biased and under estimating the variance although the impact may be limited when the proportion of truncation is small. Likelihood based methods were proposed in the literature. However, the methods can be complicated and suffered convergence problems because of the numerical integration. In this talk, we investigate a multiple imputation approach. Because the truncated values are known to be less than the detection limit, special considerations are needed to implement the imputation process. We evaluate this approach and compare it with the simple imputation and the full maximum likelihood approach with simulations. An application to a vaccine trial will also be presented.

Data Driven Method For Handling Missing Data

✦Jagannath Ghosh, Bausch and Lomb, 239 Greystone Lane, Apt # 19, Rochester, NY 14618 USA, *jagannath.ghosh@bausch.com*

Key Words: Missing Data, Pharmacokinetics, Difference Methods

There are handfuls of methods which deal with missing data. In our paper, we will present missing data problems with the idea that missing data can be handled from the available data points by using divided difference methods. After computing missing value from pharmacokinetics dataset, a relative error rate will be computed and sensitive analysis will be performed.

To Model Or Not To Model In Regression With Missing Covariates

✦Nanhua Zhang, University of Michigan, 1415 Washington Hgts, 4th floor, Biostatistics Department SPH, Ann Arbor, MI 48109, *nhzhang@umich.edu*; Rod Little, University of Michigan

Key Words: Complete-case analysis, Ignorable likelihood, Nonignorable modeling, Outcome dependency

We consider regression with missing covariates in this paper. Common methods include: (1) Complete-case analysis (CC), which discards the incomplete cases; (2) Ignorable likelihood methods (IL), which base inference on the observed likelihood given a model for the variables, without modeling the missing data mechanism; (3) Nonignorable modeling (NIM), which bases inference on the joint distribution of variables and the missing data indicators. CC and IL methods do not model the missing data mechanism while NIM models the joint distribution of variables and the missing data indicators. In this paper, we study the effect of covariate missingness on the estimation of the regression and answer the question when it is necessary to model the missing data mechanism. We will study two aspects of covariate missingness on the estimation of regression: (1) nonignorability, which concerns mainly how IL methods perform under varying levels of nonignorability; (2) outcome dependency, which studies the relatedness of covariate missingness to the outcome on the estimation of regression. We compare different methods for regression with missing covariates using a series of simulation experiments.

Methods For Handling Missing Data In Clinical Trials

✦Di An, Merck & Co., Inc., , *di_an@merck.com*

Key Words: missing data, clinical trials, multiple imputation, longitudinal study

Missing data raise a very common issue in clinical studies. Specifically in studies involving longitudial data, occurance of missingness is almost inevitable. Common missing values could be caused by reasons such as skipped visit, inadequate sample, or premature discontinuation from the treatment. This research will focus on various types of missingness in clinical trials, including intermittent missing and monotone missing values. Different approaches to handle missing data are discussed and compared.

Analysis Of Zero-Inflated And Over-Dispersed Count Data With Missing Values

✦Huiling Li, Sanofi-Aventis, 200 Crossing Blvd, P.O.Box 6890, Bridgewater, NJ 08807, *huiling.li@sanofi-aventis.com*; Lynn Wei, Sanofi-Aventis; Hui Quan, Sanofi-Aventis

Key Words: Missing data, Negative binomial regression, Overdispersion, Poisson regression, Zero-inflated

Count data in clinical research are commonly characterized by overdispersion and excess zeros. In many instances, the data can be censored if patients withdraw before the study ends. Poisson regression model based on the assumptions of equi-dispersion and non-informative missing data might not provide appropriate statistical inference to the real-life count data. In our study, this method is compared with Poisson regression with generalized estimating equations (GEE), negative binomial model, zero-inflated Poisson model (ZIP), and zero-inflated negative binomial (ZINB) model under different simulation scenarios. The comparison is performed in terms of type I error, power, and the impact of the informative and non-informative missing data to evaluate the robustness of the models. A data example is presented to illustrate the application of these analysis approaches.

Testing Of Association For 1:M Matched Case Control Pharmacogenomic Study With Missing Data

✦Shuyan Wan, Merck Research Laboratories, 126 E. Lincoln Ave, RY 34A-316, Rahway, NJ 07065, *shuyanwan@gmail.com*; Peggy H Wong, Merck

Key Words: Pharmacogenomics, Missing data, Matched case control, Power

As one type of pharmacogenomic study, retrospective case control studies are often run to either explore the association of safety endpoints with potential genomic biomarkers or to predispose the better responders by genomic biomarkers to enhance drug use with potential medicine personalization. Once a case is defined, depending on the sample size of the cases, researchers may choose to use a 1:M (usually 2 = M = 4) matched case/control design to improve statistical efficiency. This paper addresses the statistical issues associated with such design in terms of power calculation and testing in both complete and missing data scenarios. Simulation results will be shown to evaluate the performance of the modified binomial products as well as conditional logistic regression approaches. A real data application will also be presented.

240 Issues with Analysis of QT/QC and Safety ■

Biopharmaceutical Section

Monday, August 1, 2:00 p.m.–3:50 p.m.

Evaluation Of Qt Prolongation In The Absence Of Placebo Group: An Exposure-Response Modeling Approach

✦Suman K Sen`, Novartis Pharmaceuticals Corporation, 180 Park Ave, Bldg 104 Room 2K14, Florham Park, NJ 07932 USA, *sumanksen@gmail.com*; Venkat Sethuraman, Novartis Pharmaceuticals Corporation

Key Words: Thorough QT study, ICH E14, exposure-response modeling

QT prolongation is an important safety biomarker in clinical development. Regulatory agencies as well as international committees (such as ICH E14) recommend thorough QT studies be done for all new chemical entities. However, for certain types of therapies (for example, oncology drugs), placebo group is not available, which requires the development of new methods to evaluate risk of QT prolongation. We address this issue taking a exposure-response modeling approach. We explore the effect of number of ECG timepoints, type of concentration-response profile, the assumed time-concentration profile and

sample size. We then present the results comparing our approach with the traditional approach of intersection-union test in the presence of a placebo group.

Nonparametric Approaches To Assess The Risk Of Qt Prolongation Of An Investigational New Drug

✦Kristin Ashley Linn, North Carolina State University Department of Statistics, 103 Bromfield Way, Cary, NC 27519, *kalinn@ncsu.edu*; Sujit Kumar Ghosh, North Carolina State University

Key Words: Bayesian inference, Monte Carlo simulation, QT interval, QT prolongation, Linear mixed models, Nonparametric

The QT interval is the portion of the heart's electrical cycle that corresponds to ventricular depolarization and subsequent repolarization. Some non-antiarrhythmic drugs may cause QT interval prolongation, which is associated with a potentially fatal condition called torsades de pointes. In this paper, we use a Gaussian process framework to estimate the maximum difference from baseline of the corrected QT interval (QTc). The mean of the Gaussian process is approximated by a sequence of Bernstein polynomials, which is shown to result in a simple linear mixed model framework. The maximum difference between the mean functions is then estimated to characterize the effect of treatment on the QTc interval. The results of a simulation study demonstrate that our proposed method has several advantages over existing methods including decreased bias and mean squared error. We apply our method to a real data set obtained from a 'thorough QT/QTc' study conducted by GlaxoSmithKline.

Estimating the Hazard Rate Accounting for Bias in Safety Reporting in a Long-Term Safety Follow-Up Study

✦Prasheen K Agarwal, Johnson and Johnson, 965 Chesterbrook Blvd., Wayne, PA 19087, *pagarwa5@its.jnj.com*; Jiandong Lu, Johnson and Johnson

Key Words: hazard ratio, reporting bias, missing data

Introduction: A phase 2 trial and a long-term safety follow-up (LTSFU) study in COPD subjects were conducted. The data beyond the primary study suffers from potential reporting bias due to spontaneous reporting during the gap (period between end of phase2 and start of LTSFU) and unblinding prior to the LTSFU study. Method: Assuming the time to onset of malignancy is exponentially distributed with true rate parameters ?pl and ?act for placebo and active groups, respectively, we model the effect of reporting bias by a random variable zi (?pl = ?pl* zi , ?act = ?act* zi) where, ?pl* and ?act* are observed rates. For the placebo group it is possible that ?pl > ?pl*, therefore, zi ~?Uni[d-0.05, d+0.05], where d>1. For the active group it is more likely an investigator would report the onset of a malignancy during the gap, therefore, zi ~ Uni[0.95, 1.05]. Simulated datasets with data from subjects with observed LTSFU data (107) and events generated for subjects without LTSFU (124) are used to obtain estimates for true hazard ratio. Results: With more than 15% under-reporting of events in the placebo group the estimated hazard ratio is less than 1.

Challenges In Reporting Registry Data Under Rems

✦Lynne Zhang, Lundbeck Inc, , *xlynne.z@gmail.com*; Lynne Zhang, Lundbeck Inc

Key Words: Registry, REMS

Risk Evaluation and Mitigation Strategies (REMS) may be required by the FDA to ensure that the therapeutic benefits of a drug outweigh the risks. A patient registry may be required as a component of a REMS, with mandatory enrollment in the registry of all patients and all prescribers or related healthcare professionals who intend to use the drug. A patient registry helps assess the risk/benefit for every patient, and thus, by nature, collects a large amount of information. A specific registry will be discussed to share the challenges of data reporting and will describe the implications of having no formal monitoring and query process, having data collection from several sources, and having data transfers by a third party. A biostatistician's responsibility for this type of registry where the agency has agreed to report the data as is, such as checking key data fields, along with lessons learned will also be discussed.

Prediction For Severe Adverse Drug Events By Systems Biology And Statistical Learning

Peiling Liu, Bioinformatics, TIGP, Academia Sinica; Institute of Biomedical Informatics, NYMU, Taiwan; Liang-Chin Huang, School of Informatics, Indiana University, USA; ✦Henry Horng-Shing Lu, Institute of Statistics, National Chiao Tung University, Taiwan, Taiwan, *hslu@stat.nctu.edu.tw*

Key Words: adverse drug event (ADE), systems biology, classification

We propose an integrated approach based on systems biology and statistical learning for severe adverse drug events (ADEs) prediction in this study. This study utilizes systems biology informatics for Drugomics feature extraction of 1163 drugs based on DrugBank, HPRD PPIDB, KEGG and GODB. The FDA report system provides historical ADEs among actual patients. Among all the adverse patient responses associated with properly prescribed medicine, we choose a set of severe ADEs such as death, life threatening events and toxicity to build the prediction models. We use advanced statistical learning methods to predict the ADE incidences in the general population. In the training models, we obtain the summary accuracy of 84.18%, 85.12% and 81.08% respectively. The prediction accuracies of 10-fold cross validation for toxicity (66.72%), death (58.56%) and life-threatening (57.87%) ADEs are higher than random guess accuracy.

Handling Hierarchical Data Structure In The Safety Evaluation

✦Huanyu Chen, FDA, 10903 New Hanpshire Ave, Bldg 21 Room 3511, Silver Spring, MD 20993, *huanyu.chen@fda.hhs.gov*

Key Words: Hierarchical, Safety Evaluation, Meta analysis, Multilevel model

Hierarchical data structure is common in the safety data. For example, patient nested within site, site nested within trial, and trial nested within drug, drug nested within drug class. To avoid publication bias using Meta analysis and adjust detailed knowledge of natural history

of adverse event in the safe review, we applied multilevel models on pooled safety dataset to detect whether a class of drug is associated with increased risk of 30 day post treatment all cause mortality.

On The Relationship Between Baseline Correction Method And Covariance Structure For Analysis Of A Thorough Qt Crossover Study

✦Wenqing Li, Novartis Oncology Biometrics and Data Management, 180 Park Ave, Bldg. 104/2K15, Florham Park, NJ 07932, *jason.li@novartis.com*; Andrea Lynn Maes, Novartis Oncology Biometrics and Data Management; Michelle Quinlan, Novartis Oncology Biometrics and Data Management; Suraj Anand, Novartis Oncology Biometrics and Data Management

Key Words: Through QT (TQT), cross over, baseline correction, variance covariance, repeated measures

Thorough QT/QTc (TQT) trials are conducted to assess a drug's risk potential for prolonging the QT interval. A randomized crossover with therapeutic and supra-therapeutic doses of a test drug, placebo and an active control is a frequently used TQT study design for drugs with a short half-life. Within each period, multiple postdose ECG readings are collected over time leading to a repeated measures scenario. In addition, baseline readings at single or multiple timepoints are collected prior to treatment. Two active, yet seemingly separate, areas of statistical research for TQT crossovers concern the method of baseline correction and the specified covariance structure for repeated measures analysis. Due to the fact that the correlation among baseline-corrected QTc values is dependent upon the definition of baseline, these two research areas cannot be considered separately. Assuming reasonable correlation patterns for the marginal baseline and postdose QTc values, we illustrate how the covariance structure for baseline -corrected QTc differs under various definitions of baseline, thereby providing guidance on suitable covariance structures given any baseline specification.

241 Challenging Populations and Novel Sample Designs

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

Monday, August 1, 2:00 p.m.–3:50 p.m.

The Mali Road Project: A Case Study In Paired Sampling

✦Betty J Jackson, US Census Bureau, 7055 Gentle Shade Road, Apt. 201, Columbia, MD 21046, *betty.j.jackson@census.gov*; Randall J Parmer, U.S. Census Bureau; Saskia R Devries, U.S. Census Bureau

Key Words: paired sample design, road improvement project, household finances

The Millennium Challenge Corporation funded a development project in Mali to improve the Niono Goma Coura road. The U.S. Census Bureau has been enlisted to provide statistical support in evaluating the economic impact of this project. A road corridor in a similar area was chosen as the control to measure the economic benefits of the road improvement project. A paired sample design was developed for housing units within the respective areas being the sampling units for a survey on household finances. This paper addresses sample design aspects of the survey, including calculations of anticipated sample sizes, power, cluster sizes, village and household pairing, sort, and stratification. A particular focus of the paper is the contrast of designing and conducting surveys in developing countries such as Mali with surveys in the U.S.

Comparing Measure-Level Sampling And Bed-Level Sampling For Group Quarters Sample Redesign

✦Stephanie M. Coffey, United States Census Bureau, 4600 Silver Hill Road, Suitland, MD 20233, *stephanie.coffey@census.gov*; Bonnie Coggins, United States Census Bureau; Arielle Gerstein, United States Census Bureau; Rachel Horwitz, United States Census Bureau

Key Words: Sampling, Clustering, Demographic Surveys

Group Quarters (GQs) are one of several types of living arrangements sampled in demographic surveys. They include college dormitories, group homes, and religious quarters. For first-stage sampling, individuals that reside in GQs are converted into Housing Unit Equivalents within a block to ensure that a unit of sample selected in a GQ corresponds to an average household (which in 2000 was 2.59 individuals). First-stage sampling is done at the PSU (groups of counties) level. Several demographic surveys are undergoing redesign to address new and continuing data needs. For this redesign, research was conducted to determine whether bed-level sampling should replace measure-level sampling by examining whether bed-level sampling would decrease the level of clustering within the sample while avoiding a significant increase in advance listing procedures in the field. This paper examines: measures, why they were used in the past and benefits and drawbacks; bed-level sampling, its benefits and drawbacks; simulations to compare clustering effects and comparative field effort of the two sampling methods. The result is a final recommendation to continue using measure-level sampling.

Composite Size Measures In Surveys Of Rare Or Hard-To-Reach Populations

✦Frank Potter, Mathematica Policy Research, Inc, P.O. Box 2393, Princeton, NJ 08543, *FPotter@Mathematica-MPR.com*; Eric Grau, Mathematica Policy Research, Inc. ; John Hall, Mathematica Policy Research, Inc.

Key Words: hard-to-reach populations, rare populations, composite size measure, complex survey design

Rare and hard-to-reach populations pose significant challenges to the design and implementation of cost-efficient sample surveys. To find and enumerate these populations, multi-stage surveys are often used to avoid the construction of a sampling frame for the entire target population, and primary sampling units (PSUs) are selected with probability proportional to a size measure related to the population sizes in the PSUs. When multiple populations are of interest, composite size measures are used that are based on the population counts in the PSUs. Some composite size measures were described by Folsom, Potter and Williams (1987) and by Fahimi and Judkins (1991). The purpose of this paper is to discuss the advantages and disadvantages of these methods for various study populations and when to use these algorithms.

We will demonstrate these size measures in surveys of students with disabilities, of persons receiving unemployment insurance compensation, and of persons in households receiving Supplemental Nutrition Assistance Program (SNAP, formerly called the Food Stamps program) payments and low income households not receiving SNAP payments.

Collecting Health And Safety Information In A Mobile And Hard-To-Reach Population: Survey Approaches For U.S. Truck Drivers

✦Karl Sieber, CDC/NIOSH, 4676 Columbia Parkway, Cincinnati, OH 45255, *wks1@cdc.gov*

Key Words: transportation, survey, truck driver

Improving public health and safety by reducing transportation-related injuries and illnesses is one goal included in the U.S. Department of Transportation's 2010-2015 Strategic Plan. To measure progress toward this goal, baseline data to determine prevalence of health and safety conditions is needed. Collecting such information in a highly mobile and difficult-to-reach population such as truck drivers requires special approaches. The type of truck driver-long-haul (over the road), regional, or local- is important since each type faces distinct problems depending on the organization of work. Data collection methods must be tailored to the type of driver. Approaches to data collection are further complicated due to factors such as the mobile nature of the population and lack of a suitable sampling frame. For example, only about 10% of drivers are registered union members. Some long-haul drivers may drive for weeks before arriving home, or may consider their truck as their home. Various approaches to data collection have been used in these types of population. Approaches to surveying truck drivers are presented as well as lessons learned in dealing with this population.

Using Ancillary Information To Stratify And Target Young Adults And Hispanics In National Abs Samples

Charles DiSogra, Knowledge Networks, Inc; ✦J. Michael Dennis, Knowledge Networks, Inc., , *mdennis@knowledgenetworks.com*; Erlina Hendarwan, Knowledge Networks, Inc.

Key Words: ABS, mail survey, targeting, stratification, ancillary data

The USPS DSF frame is used by Knowledge Networks to recruit members for its probability-based web panel. A bilingual invitation packet with $2 is mailed; follow-up mailings go to non-responders. Households without Internet access are provided a laptop and free ISP for participation. In 2010, four large (45,000) samples were fielded in two waves each. This design targeted Hispanics in one stratum using census block (CB) data. The other stratum had the balance of CBs. Simultaneously, ancillary data from commercial databases attached to each address were tracked as potential predictor variables for future targeting purposes. Hispanic ethnicity and age were key items. Based on the 2010 findings, samples for 2011 were designed using only ancillary data to make four strata: Hispanic 18-24, all else 18-24, Hispanic 25+, all else 25+. Mailing materials/incentives were unchanged. So far, 2 surveys of 45,000 addresses each were fielded with this design. Preliminary yields compared to the CB design show the yield of young adults improved by >10% and the yield of Hispanics by >12%. Results on the efficiency of an ancillary data stratification design for ABS samples will be shown.

Two Methods Of Sampling New Construction For The Demographic Surveys Sample Redesign

✦Andreea Rawlings, U.S. Census Bureau, 4600 Silver Hill Rd, Washington, DC 20233, *andreea.m.rawlings@census.gov*; Danielle Castelo, U.S. Census Bureau

Key Words: skeleton sampling, new construction sampling, Master Address File

Historically for the demographic surveys, we selected a decade worth of sample after each decennial census and divided it into quarterly or monthly samples for interviewing. This raises the issue of how to capture housing units which are built after the sample is selected. In the past, these new construction units were captured by selecting "placeholder" units which were later matched to actual growth to determine which units would be in sample. We refer to this method as skeleton sampling. As we move to sampling from the Master Address File, we have an alternative method available, which is to directly sample new construction; this method of sampling mimics the sampling of the original units. In this paper we provide an overview of the two methods for sampling new construction, discuss the benefits and drawbacks, and evaluate each method's impact on variances and operations.

Using Order Sampling To Achieve A Fixed Sample Size After Nonresponse

Pedro J. Saavedra, ICF-Macro; ✦Lee Harding, ICF-Macro, 11785 Beltsville Drive, Suite 300, Calverton, MD 20705, *lHarding@icfi.com*; Francine Barrington, ICF-Macro

Key Words: replacements, Pareto sampling, adjustments, simulations, sequential Poisson sampling, propensity categories

There are situations when a study requires a fixed sample size, either for contractual reasons or because the cost of collecting data for too many cases is prohibitive. This makes the preferred practice of oversampling and then adjusting for nonresponse impractical. Under certain conditions a simple random sample can be obtained by randomly sorting the frame and selecting the first n in the random order. This yields a fixed initial sample size, but a variable respondent sample. In a case where potential respondents beyond the targeted number of completes can be approached in sequential order (exhausting contact attempts before going to the next unit), the sampling process can continue until the desired number of completes is obtained. Nonresponse adjustments can then be made as if the combined set of respondents and nonrespondents constituted an initial sample. A similar approach to the one described above could be used to achieve a fixed number of completes using Sequential Poisson Sampling or Pareto Sampling. Here the probability of selection is changed, but the difference may be minimal. Simulations using SRS, SPS and Pareto were conducted to examine this practice.

242 Multiple Modes and Multiple Data Sources ■

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

Monday, August 1, 2:00 p.m.–3:50 p.m.

A Randomized Experiment Comparing Response Quality In A Single And A Mixed Mode Design Health Survey

✦Johann Carstensen, University of Erlangen-Nuremberg, Findelgasse 7/9, Nuremberg, International 90402 Germany, *johann.carstensen@wiso.uni-erlangen.de*; Peter Kriwy, University of Erlangen-Nuremberg; Gerhard Krug, University of Erlangen-Nuremberg

Key Words: survey design, mixed mode, sensitive items, web survey, nonresponse, social desirability

Asking sensitive questions in surveys is prone to mode effects. In a randomized experiment we compare a single mode CATI to a mixed mode survey where respondents can chose between a CATI and Web-based mode. Our main focus lies on the consequences of the additional Web-survey option on data quality. Indicators of data quality are the respondent`s tendency towards social desirable answers, satisficing behavior and item non-response. In both modes identical questionnaires are used covering sensitive items such as health behavior and self-reported illness and including auxiliary variables e.g. about item sensitivity and affinity to the internet. However differences in means of sensitive items between single and mixed mode surveys can also be due to differences in unit non-response induced by the Web-option. Therefore we propose a decomposition method. It utilizes statistical matching techniques to estimate even without external validation data the relative strength of selection and mode effects. As the data collection will start in April 2011 we will be able to provide preliminary results by July. The project is funded by the German Robert Koch Institute (FKZ 1362/1-978).

Does Survey Mode Make Differences? - A Comparative Evaluation Of The Department Of Defense Survey Of Health Related Behaviors And The Us Air Force Community Assessment Survey

✦Zhiwei Zhang, ICF International, 9300 Lee Highway, Fairfax, VA 22031, *zzhang@icfi.com*

Key Words: military survey, mixed-mode, web survey, substance use, mental health, DOD

For three decades, the US Department of Defense (USDOD) has collected health related information through the Survey of Health Related Behaviors (HRB)among Military Personnel. For two decades, the US Air Force (USAF) has conducted the Air Force Community Assessment (AFCA)Survey. Through both surveys, estimates of a variety of behavioral and health related outcomes have been obtained to support top level policy makings and assess military readiness. These estimates have been compared across different service types and between military and civilian populations, but seldom across different surveys in different data collection modes for the same military populations at the same time. This study examines the extents that alcohol and drug use and suicidality prevalence rates may differ in different surveys and whether these possible differences are systematic, using the survey data of the 28,000 service members (including 7,000 air force active duty personnel) across 64 military installations worldwide collected primarily through on-site group sessions and from 80,000 Air Force active duty members across all major air force commands globally collected via web survey.

Comparing Cell Phone And Web For A Student Survey

✦Young-Je Woo, Dongguk University, 26-3, Pil-dong, Jung-gu, Seoul, 100-715 South Korea, *youngjw@dongguk.edu*; Sun-Woong Kim, Dongguk University; Mick Couper, University of Michigan

Key Words: cell phone survey, web survey, coverage, response rate

In many countries the proportion of the population with cell phones is higher than that with Internet access. This is particularly true of the college population. Given the recent attention focused on surveys of cell phone users, and the preference for using the Internet to survey college students, it is useful to compare these two modes of data collection. This paper reports on a mode experiment conducted in the 2010 Time Use Survey of students at Dongguk University in South Korea. The frame of registered students contains both cell phone numbers and e-mail address. A sample of students was selected from the list and randomly assigned to a CATI or Web survey mode. Students were notified of the survey in both modes using both text messaging and e-mail messages. Findings show that the cell phone survey has a distinct advantage over the Web survey concerning response rates, coverage of domains, and item nonresponse. Substantive differences between the two modes were found for about half the survey questions. This suggests that cell phone surveys may be useful to surveys in populations with universal or near-universal coverage, and where cell use may be more popular than Internet use.

Respondent Effects In A Dual-Mode Survey Of Physicians: 2008-2009

✦Esther Hing, National Center for Health Statistics, 3311 Toledo Road, Room 3409, Hyattsville, MD 20782, *ehing@cdc.gov*; Chun-Ju Hsiao, National Center for Health Statistics; Paul Beatty, National Center for Health Statistics; Sandra Decker, National Center for Health Statistics

Key Words: Physician survey, Dual mode survey, Respondent effects

The National Ambulatory Medical Care Survey (NAMCS) is an annual survey of office-based physicians and visits to their practices. Since 2008, the original physician sample was augmented with a supplemental sample of physicians who responded to NAMCS questions on electronic medical record (EMR)/ electronic health record (EHR) systems through a mail questionnaire. Mail and face-to-face survey data from 2008 and 2009 were combined to produce dual-mode estimates. This paper compares how survey respondent (physician or office staff members) and mode affected responses to questions about EMR use in the 2008 and 2009 surveys. In 2009, but not in 2008, reports of overall EMR/EHR system use were associated with respondent type and survey mode, after controlling for quarter of year, practice size, specialty category, and metropolitan area status. Response to having basic and fully functional systems did not vary by respondent type or survey mode. In both years, the proportion of physician respondents was higher in the mail survey (52% in 2008 and 62% in 2009) than in personal interviews (26% in 2008 and 10% in 2009).

When Survey Data Come From Multiple Sources

✦James R Chromy, RTI International, PO Box 12194, 3040 Cornwallis Road, Research Triangle Park, NC 17709, *jrc@rti.org*

Key Words: Administrative Data, Missing Data, Unit Nonresponse, Item Nonresponse

A challenging situation arises when survey data comes from more than one respondent source (e.g., student, parent, or teacher) and administrative data are used to supplement sample survey data. In some education surveys, administrative data can be available from the institution or from external data bases. If the same information can be obtained from more than one source, some rules must be developed to decide which source takes precedence if disagreement occurs. Data missing from one source might be imputed with data from an alternate source. An especially interesting situation arises in defining unit and item response rates. A useable case rule must be developed to identify key information items or combinations of these items to qualify as a unit respondent. The useable case rule can sometimes be satisfied with data from only a subset of the sources. Equally challenging are the methodologies for adjusting for unit and item nonresponse in the analytic files produced from the survey. This paper poses some hypothetical situations and discusses tradeoffs between unit respondent sample size and data record completeness. This comparison is also examined in terms of sampling er

Missing Data In Record-Linked Data Sets. Comparing The Performance Of Different Missing Data Techniques

✦Gerhard Krug, University of Erlangen-Nuremberg, Findelgasse 7/9, Nuremberg, International 90402 Germany, *gerhard.krug@wiso.uni-erlangen.de*

Key Words: missing data, record linkage, multiple imputation, sample selection model

Combining data from a survey with register data using record linkage (RL) can lead to missing data and potentially to biased estimates, if survey respondents have to consent to it. Missing data (MD) techniques can be used to correct for potential record linkage bias. Based upon a survey where participants were asked permission for RL the performance of different missing data techniques is compared. For respondents who refuse their permission I set their survey answers to missing, creating pseudo-missing data. To correct for potential bias, OLS Regression is performed using complete case analysis (MCAR), multiple imputation (MAR) and Heckman's sample selection model (MNAR), respectively. Their performance is compared to a benchmark regression that is based on the complete data set. Several missing data scenarios are compared. Results indicate that when RL-bias was small, all missing data techniques performed well. In contrast, when RL-bias was high, only multiple imputation was able to correct for the RL-bias, given that only independent variables had missing values. With high RL-bias and missing values in the dependent variable, none of the MD techniques eliminated the bias.

An Integrated Adaptive Approach To Data Fusion

✦Hui Xie, University of Illinois, , *huixie@uic.edu*; YI Qian, Northwestern University

Key Words: Data Combination, Nonparametric Method, MCMC, Survey, Multiple Imputation, Missing Data

Data fusion combines data items from various sources based on a common set of variables. Using the synthesized database, researchers can overcome the limitations of a single-source dataset and answer important questions that cannot be addressed otherwise. We propose an integrated adaptive imputation approach to data fusion method. The proposed method can handle a mixture of continuous, semicontinuous and discrete variables in a robust manner in that no parametric distributional assumption is required for any variable in the data. Therefore the method is applicable to any distributional shapes and can adaptively and automatically generate suitable distributions for any variables to be fused. A simulation study is conducted and shows superior performance of the method as compared with prior approaches. We then apply it to a survey study on counterfeit. The analysis demonstrates that the proposed method can increase the efficiency and validity of data fusion and make data fusion more powerful.

243 Teaching Stats in Health Sciences

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

Monday, August 1, 2:00 p.m.–3:50 p.m.

Higher Order Assessments: Bridging The Gap Between Expectations And Outcomes

✦Heather M Bush, University of Kentucky College of Public Health, 121 Washington Ave, Rm 205C, Lexington, KY 40536-0003, *heather.bush@uky.edu*

Key Words: teaching, assessment, student outcomes, application, thinking skills, evaluation

Recently, many statistics courses in the health sciences have been restructured to focus on concepts and applications with the intent that students would be better prepared consumers and producers of statistical information. However, even with a shift in focus, a gap still exists between instructor/student expectations and student outcomes. Specifically, after completion of course(s), students continue to struggle with applying statistical reasoning/thinking to research problems. One explanation for this is that the method of assessment has not been updated with course content; many assignments still focus on memorization and repetition. Using Bloom's taxonomy for thinking skills, it is possible to explain the gap between objectives and outcomes since expectations involve higher order thinking skills while assignments and evaluation focus on lower order thinking skills. One way to bridge the gap between objectives and outcomes is to construct assignments that provide students opportunities to employ higher order thinking. Examples of implementing higher-order assignments will be discussed. In particular, using published articles as a mode of assessment will be explored.

A Course In Medical Statistics: Comparing Team-Based Learning Strategies With Traditional Methods

✦Sandra Stinnett, Duke University Medical Center, , *stinn001@mc.duke.edu*; Colleen Grochowski, Duke University Medical Center

Key Words: medical statistics, team-based learning

At Duke University School of Medicine, third-year medical students carry out research with a mentor for a year, analyze their data and write a thesis on their research. The medical statistics course is offered at the beginning of the third year to prepare the students for their research. Two cohorts of students completed the course in 2010: a cohort of 35 students, using team based learning (TBL) and a cohort of 14, using online materials and web-based exams. The TBL class met for 2.5 hours 2-3 times per week for 4 weeks for a total of 9 TBL sessions. Each session included an Individual Readiness Aptitude Test, Group Readiness Aptitude Test, and Group Application Exercises. The Group Application Exercises consisted of 7-10 problems that involved thought and analysis of data using JMP software. Scores on all of these activities were the basis for the students' grade in the course. The online course consists of 9 sets of PowerPoints/streaming videos and 4 exams. The TBL cohort reported that the class was enjoyable, engaging and effective. Away-students did well in the course, but reported a need for more engagement. We present and contrast the TBL methodology and online methods.

Virtual Discussion For Real Understanding

✦Kendra K Schmid, University of Nebraska Medical Center, 984375 Nebraska Medical Center, Omaha, NE 68198-4375, *kkschmid@unmc.edu*

Key Words: distance learning, online, discussion

One of the challenges of teaching is engaging students in a subject they do not see as relevant to them. This issue is especially prevalent when teaching statistics to health science students as many do not consider statistics an important piece of their medical training. Additional difficulty is presented when teaching courses via distance technology or courses that are partially or completely online as the valuable class discussion component is lost. This paper focuses on fostering "discussion" about statistical concepts and how they relate to each student on an individual level. The approach includes an online discussion board where students participate in guided questions and post and critique an article related to their field of study. The objectives are to enhance knowledge, develop critical thinking, and gain an appreciation of how statistics is used in their field. Students must reflect on why statistics is important in their field and respond to other students' posts. This approach has been successfully piloted in an online class for Allied Health students and is currently being used in a graduate level class including both synchronous and asynchronous distance learners.

Some Strategiesthat Contribute To The Mastery Of Statistical Methods

✦Winston Ashton Richards, Penn State Harrisburg, 777 W. Harrisburg Pike, Middletown, PA 17057, *ugu@psu.edu*

Key Words: Statistics Students, Teaching, Basic Statistics

Some Strategies That Contribute To the Mastery of Basic Statistical Methodology. Students enroll in my Statistics course either because they are required to take my course or they have a strong interest in Statistics. As an instructor therefore I am faced with two initial challenges. I have to capture the interest of those required to take my course by presenting the topics in such a way that students will see the relevance, importance, pervasiveness and beauty of Statistics. The hope is that they will become like the group who are taking my course by choice. With the second group the task is less formidable because they come with an intense desire to master the subject matter, and I proceed in a manner that will sustain their interest and enable them to further develop their statistical skills. This presentation will consist of the strategies I have employed to accomplish these goals as a faculty member at Penn State Harrisburg, at The University of The West Indies and at Stanford University.

Making Data Talk: Communicating Health Data With Lay Audiences

✦David E. Nelson, National Cancer Institute, NIH, 6120 Executive Blvd., Ste 150E, MSC 7105, Bethesda, MD 20892-7105, *nelsonde@mail.nih.gov*; Brad Hesse, National Cancer Institute, NIH; Harry T. Kwon, National Cancer Institute, NIH

Key Words: health data communication, data selection, data presentation

The ability to understand statistical data is critical in order to raise awareness and make key informed decisions. Statistical data are helpful because they can provide the rationale behind scientific understanding and health recommendations. However, communicating health data, especially to lay audiences, is difficult because of low levels of mathematical and scientific literacy, cognitive limitations, existing lay beliefs, desire for certainty from experts, and emotions. Based on a comprehensive review of the scientific and practice literature, this presentation will discuss research on the selection and presentation of health data and provide practical suggestions on how statisticians can better communicate data to lay audiences. This session will introduce the OPT-In (Organize, Plan, Test, Integrate) framework which can assist with communication planning and decision making when selecting and presenting data to lay audiences. Case studies utilizing this framework to discuss cancer control/prevention efforts will be presented to demonstrate the important role that selecting and presenting data can play in raising awareness of important public health issues among lay audiences.

On Calculating P-Values For Discrete Distributions

✦Hui Xu, St. Cloud State University, 720 4th Ave S, ECC 250, St. Cloud, MN 56301, *hxu@stcloudstate.edu*

Key Words: P-values, Binomial Distribution

It is necessary to calculate P-values in various situations in order to teach the well known controversy about P-values that they do not always provide objective evidence against null hypotheses. However, in introductory statistics courses, most of the P-values are computed based on normal, t and other continuous distributions or through software packages. In this talk, I will use some examples to illustrate the procedure of computing P-values in discrete distributions. Hopefully, this will give students a better understanding about P-values.

244 International Social Statistics

Social Statistics Section, Section on Government Statistics, Statistics Without Borders, Scientific and Public Affairs Advisory Committee

Monday, August 1, 2:00 p.m.–3:50 p.m.

Statistical Control Charts For Detecting And Correcting The Under-Reporting Of Adult Mortality In Latin American Countries.

✦Natalia Rojas-Perilla, Universidad Nacional de Colombia, Bogot·, 1 Colombia, *nerojasp@unal.edu.co*

Key Words: Adult mortality, Under-reporting, Generalized Grow Balance method (GGB), Control charts for individuals and attributes

This paper proposes the use of quality control charts to analyze the process of adult mortality in Colombia during the twentieth century, to identify and provide valuable information to carry out a correction of under-reporting of mortality by the method 'Generalized Grow Balance' proposed by Hill(1987) in which underreporting coverage is estimated by the growth and mortality rates between censuses. For the detection and correction, mortality data used were annual and adult, disaggregated by sex from 1917 to 2007 and the census of 1912, 1918, 1938, 1951, 1964, 1973, 1985, 1993 and 2005. Were adjusted time series models for both men and women, in order to eliminate non-uniformity in the data and to make appropriate use of control charts for individuals observations. Was found that control charts were able to corroborate the historical facts themselves of underreporting of mortality in Colombia during the twentieth century, therefore, have the same for the monitoring of mortality in countries like Colombia, where the under-reporting has been an implicit default in the registers of vital statistics, certainly would detect significant changes in the population dynamics of the country.

A Demographic Study of Freemasons in Cuba: 1945-2008

✦Jorge Luis Romeu, Syracuse University, Dept. Mech & Aerosp. Eng., P. O. Box 6134, Syracuse, NY 13217, *jlromeu@syr.edu*

Key Words: demographic study, population at risk, time series, sociopolitical study

Cuban Freemasons constitute one of the oldest, most numerous and most geographically widely spread, organization of the Cuban civil society. Its characteristics, including its religious and political tolerance and open-minded philosophy, can positively contribute to strengthen Cuban civil society. In this paper we examine the evolution of Freemasons in Cuba, using annual membership data from the Grand Lodge, in Havana. Defining the Masonic equivalent of population at risk, a new indicator is derived. Then, using Immigration and Naturalization Service (INS) data, estimates of the number of Freemasons that left Cuba in the wake of Castro's revolution, are obtained. Using United Nations data, estimations of deceased Masons are obtained. Using internal membership change data, bounds for number of Communist Party and Administration officers that may have joined the Freemasons, after the 1992 lift of the prohibition for belonging to this organization, are obtained. In this manner, we describe the evolution of members, through our seven defined epochs of Cuban sociopolitical developments, during the second half of the XX Century.

China'S Higher Education Expansion And The College Wage Premium : An Empirical Study Based On The Data Of Chns

✦Xu Sun, Dongbei University of Finance and Economics, No. 217 Jianshan Street, Shahekou District, Dalian, International 116025 China, *sunxu@dufe.edu.cn*

Key Words: Higher Education Expansion, college wage premium, Data of CHNS

Using the data of China Economic, Population, Nutrition, and Health Survey(CHNS), this paper reports estimates of the China "college wage premium" for young graduates (21-30 years old) from 1991 to 2006- a period when the higher education participation rate increased dramatically. Our analysis suggests, quite remarkably, that despite the large rises in higher education participation in the 1990's through to the mid 2000's, there has been no significant fall for men and even a large, but insignificant, rise for women. To explore how the expansion affected the college wage premium we present estimates, using the micro-data, by quartile of the conditional wage distribution. Quantile regression results reveal a fall in the premium only for men in the bottom quartile of the distribution of unobserved skills.

Is Poverty A Determinant Of Educational Attainment? Perspective From Africa

✦Oyelola Abdulwasiu Adegboye, American University of Afghanistan, P.O. Box 458, Central Post Office, Durulaman Road, KABUL, International 5000 Afghanistan, *aadegboye@auaf.edu.af*; Danielle Kotze, University of the Western cape

Key Words: Education, Millennium Development Goals, income, inequalities, population health, poverty

Access to education particularly in the developing countries has been discouraging. The Jomtien 1990 declaration of the World Conference on Education for all stipulates that every person (child, youth and adult) shall be able to benefit from educational opportunities designed to meet their basic needs. It stipulated that by 2015, the time set for Millennium Development Goals, all children must have access to and should be able to complete primary school. Many have described income and education as the fundamental determinant of health and as indicators for socio-economic status. Schooling improves empowerment, productivity, health, and reduces negative features of life such as child labor. Many research has been carried out on the association between income inequalities and population health and much of this work reported that rich and educated people live longer and suffers less mortality than poor and less educated people. In this research we will provide empirical studies on population studies and poverty, educational attainment and income inequalities. We will also investigate the area-level distribution of education equality associated with individual income outcomes in Africa

Expenditure Pattern of Undergraduate Students in the Department of Mathematics and Statistics University of Cape Coast Ghana

✦DANIEL AGYEKUM AMAKYE, DANIEL AGYEKUM AMAKYE, C/O ENCHILL YAW MARK, MICHEL CAMP JHS, BOX 385,, MICHEL CAMP , TEMA, ACCRA, 00233 GHANA, *amakyedagyekum@yahoo.com*

This study was conducted on the students of the department of mathematics & statistics,university of cape coast in the 2008/2009 academic year.it investigate the expenditure of the students on accommodation/ utilities,feeding,school fees,learning materials,transportation, clothing and other expenditure.a sample of 208 students was selected from a population of 656 students.a primary data were collected from the respondents using self administered questionnaire and a multistage probability sampling.the data were analyzed using statistical product for service solution (spss)and microsoft excel.the study shows that a student spends on the average 13,235.52($1,352.52) in the 2008/2009 academic year.student in the age group of 30-34 spend more than any other age groups students.student who are married spend more than the single.residents student spend more than non resident students. school fees,feeding,and accommodation/utilities constitute the major component of student expenditure.there is no significant difference between male and female expenditure.on the average level 200 students spend more than those of other levels.

A Study Of Nonresponse Adjustment For Subject Missing- A Teps Case Study

✦Hong-Long Wang, National Taipei University, 151, University road,, San Shia district, New Taipei City, 237 Taiwan, ROC, *hlw@gm.ntpu.edu.tw*; Ting-Hsiang Lin, National Taipei University; Lyinn Chung, National Taipei University; Shu-Yi Lei, National Taipei University

Key Words: subject missing, missing mechanism, MCAR, MAR, TEPS, weighting adjustment.

Subjects missing in a survey may be not random and cause inference bias. Non-response adjustment maybe needed to reduce the bias. However, missing mechanism will be needed before any non-response adjustment. In this paper, through a study of "Taiwan Education Panel Survey" (in short as TEPS), we will explore missing mechanism for subjects missing on 2nd wave data of TEPS. Distinguish missing completely at random (MCAR) from missing at random (MAR). To explore the variables causing MAR, we randomly select 100 data sets from the complete part. Each of the selected data set, with the sample size twice of the missing part, will combine with missing part and run logistic models to find the significant variables. Stability of the model will be discussed. For the non-response adjustments, we will apply weighting adjustment and list-wise delection and compare the result with the baseline.

Exploring New Models for Population Prediction in Detecting Demographic Phase Change for Sparse Census Data

✦Arindam Gupta, Burdwan University, Department of Statistics, Burdwan University, Golapbag, Burdwan, International 713104 India, *stat_agupta@buruniv.ac.in*

Key Words: Population Prediction, Relative Growth Rate, Demographic Phase Change

Logistic model has some limitations when applied for developing countries. In such situation the relative growth rates (RGR) exhibit some unusual trends (increasing, primary increasing and then decreasing). To tackle those situations we extend the logistic law by incorporating nonlinear positive and negative feed back terms. Here we have assumed that RGR is a function of size and time separately. The time covariate model has some key advantages than the size covariate model. It can detect the demographic phase change point at which a developing country switches over towards a developed one.

245 Models for Longitudinal and Clustered Data ■●

ENAR, International Indian Statistical Association

Monday, August 1, 2:00 p.m.–3:50 p.m.

On Mediation Analysis With Terminal Measures Data

✦Roshan N. Serasinghe, Kansas State University, Department of Statistics, 101 Dickens Hall, Manhattan, KS 66506, *roshan@ksu.edu*; Gary L. Gadbury, Kansas State University; Jeffrey M. Albert, Case Western Reserve University ; Mark Beasley, University of Alabama at Birmingham; Brad A. Rikke, University of Colorado at Boulder ; Thomas E. Johnson, University of Colorado at Boulder ; David B. Allison, University of Alabama at Birmingham

Key Words: Causality, Indirect effect, potential outcomes

The total causal effect of a treatment on an outcome variable, Y, may include the direct effect of a treatment on Y and an indirect effect through another variable(s), say Z. This indirect effect is called a mediating effect (ME), and the analysis of such data is mediation analysis (MA). Situations can arise where Z and Y cannot both be measured on an individual unit. Mouse and plant experiments are two examples where measurement of Z requires terminating the animal or plant and Y is to be measured at a later time. We refer to such data as terminal measures data. Another situation may be where one experiment focused on measurement of Z, and the second on measurement of Y, and interest is in combining the data sets to evaluate the ME of Z on Y. This talk discusses the issues involved in analyzing terminal measures data. An imputation approach is proposed to construct an individual-level mediation model that relies on a blocking variable. The technique is illustrated on a mouse data set from two experiments, one where Z was measured and another where Y was measured. Sensitivity of results to a nonestimable partial correlation between Z and Y are discussed.

Comparing Dose-Time-Response Models

Ramu Goud Sudhagoni, South Dakota State University;
✦Gemechis Djira, South Dakota State University, SHH 123, Department of Mathematics and Statistics, Brookings, SD 57007, *Gemechis.Djira@sdstate.edu*; Frank Bretz, Novartis Pharma AG

Key Words: Dose-time-response, Dynamic model, Target dose, repeated mesures

Understanding and characterizing the dose-response relationship is always a difficult step in dose-response studies. The dose-response relationship describes the change in effect on subjects caused by different dose levels after certain exposure time. In the literature, many linear and non-linear models (e.g., linear, emax, and exponential) exist for modeling dose-response data. When responses are recorded at more than one time point on the same subject, we use dose-time-response modeling. The purpose of this research is to explore existing methods such as longitudinal models and dynamic models and compare them in

terms of their efficiency. That is, identifying the overall dose-response signal and estimating the probability of target dose and precision of estimation.

Random Effects Coefficient Of Determination For Mixed And Meta-Analysis Models

✦Eugene Demidenko, Dartmouth College, Hanover, NH 03756, *eugened@dartmouth.edu*

Key Words: clustered data, random effects, growth curve, dummy variable

The key feature of a mixed model is the presence of random effects. We have developed a coefficient, called the random effects coefficient of determination that estimates the proportion of the conditional variance of the dependent variable explained by random effects. This coefficient takes values from 0 to 1 and indicates how strong the random effects are. If this coefficient is close to 0, there is weak support for random effects in the model because the reduction of the variance of the dependent variable due to random effects is small; consequently, random effects may be ignored and the model simplifies to standard linear regression. The value of this coefficient apart from 0 indicates the evidence of the variance reduction in support of the mixed model. If random effects coefficient of determination is close to 1 the variance of random effects is very large and random effects turn into free fixed effects---the model can be estimated using the dummy variable approach. Coefficient of determination is illustrated by three examples of mixed and meta-analysis models.

(In)Consistency Of Maximum Likelihood Estimators In Ornstein-Uhlenbeck Autocorrelation Tree Models

✦Cecile Ane, University of Wisconsin, Dept. of Statistics, Univ of Wisconsin-Madison, 1300 University Ave., Madison, WI 53706, *ane@stat.wisc.edu*; Lam Ho, University of Wisconsin

Key Words: tree autocorrelation, dependence, microergodicity, Ornstein-Uhlenbeck, phylogenetics, evolution

We consider linear models with strong hierarchical autocorrelation, and prove the consistency and asymptotic normality -or lack thereof- of maximum likelihood estimators. Observations are modeled at the tips of a tree and assumed to arise from a Brownian motion or an Ornstein-Uhlenbeck process along the tree. The autocorrelation between two tips increases with the length of their shared path from the root. These models are most often applied in evolutionary biology, where different tips represent different species. We show that the maximum likelihood estimators of some parameters are not consistent in standard asymptotic frameworks. In fact, these parameters are not microergodic: no estimator can ever be consistent for such parameters. We will show the analogy and differences with Ornstein-Uhlenbeck models in spatial statistics. For microergodic parameters, we will present consistency and asymptotic normality of their maximum likelihood estimators under a regular asymptotic framework. We will discuss the consequences of these results for sampling design, in application to evolutionary comparative studies.

External Validation Of Nomograms In The Presence Of Missing Data

✦Sujata Patil, Memorial Sloan-Kettering Cancer Center, 307 E. 63rd, third floor, New York, NY 10065 USA, *patils@mskcc.org*

Key Words: nomograms, predictive model, data analysis, simulation study

Nomograms are graphical depictions of multivariate models. Clinicians use these statistical tools to provide a patient-specific overall probability of an outcome based on clinical and pathological characteristics. Nomograms are often built on single institutional datasets and internally validated by dividing the dataset into training and test groups or by bootstrapping. The concordance index (CI) is one statistic that describes the discriminative ability of the nomogram. External validation is critical before a nomogram is accepted as a useful predictive tool. Valuable opportunities to externally validate a nomogram are rare and often do not come to fruition because either external datasets do not contain a variable or there is limited information on several variables. One approach to deal with this issue is to use imputation. Here, the effect of imputation on the magnitude and variability of the CI calculated in an external test dataset is assessed. The results from simulation studies done to identify the effect of imputation on the CI under several scenarios will be reported. Suggestions on useful characteristics of an externally validating dataset will be discussed.

Tests Of Separate Hypotheses For The Linear Mixed Model

✦Che Smith, UNC Chapel Hill, 3101 McGavran-Greenberg, Campus Box 7420, Chapel Hill, NC 27599-7420, *ches@email.unc.edu*; Lloyd J Edwards, University of North Carolina at Chapel Hill

Key Words: bootstrap, information criterion, model selection, nonnested models, statistical computing

In model selection, comparisons between linear mixed effects models with non nested fixed effects have been underexplored. We compare the performance of a Cox Test of Separate Hypotheses with the Extended Information Criterion to select between two candidate repeated measurement models arising from both real and simulated data. Other cases of non nested models are addressed, including the comparison of linear vs. nonlinear mixed effects models.

The Intracluster And Intercluster Correlations In Cluster Randomized Trials

✦Robert E. Johnson, Virginia Commonwealth University, Department of Biostatistics, 730 E. Broad Street, Richmond, VA 23298-0032, *rjohnson@vcu.edu*; Tina Duong Cunningham, University of California Los Angeles

Key Words: cluster randomized trials, intracluster correlation, ICC, induced correlation, finite population correction

The intracluster correlation (ICC) plays a significant role in the planning and analysis of cluster randomized studies. Two measures sampled from the same randomly chosen cluster will be correlated. It is key that the clusters are randomly selected. If fixed-e.g., a convenience sample of clusters is used-then no correlation between intracluster measures is induced. However, if the-nonrandomly chosen-clusters are randomly allocated to treatment arms, then not only is an intracluster correla-

tion induced, but an intercluster correlation between measures in different clusters is induced. The variance of contrast in treatment means is less than the variance when randomly chosen clusters are randomly allocated. The reduction in variance is related to the finite population correction employed in sample surveys. This correction is negligible when the number of clusters is large, but may play a role in sample size determination and analysis with few clusters. We will illustrate how the sampling method and randomization may induce correlation and discuss the impact on analysis.

246 Seasonal Adjustment ■

Business and Economic Statistics Section
Monday, August 1, 2:00 p.m.–3:50 p.m.

A Graphical Test Of The Quality Of Seasonal Adjustment: An Empirical Comparison Of Various Methods Of Seasonal Adjustment

✦Raj K Jain, BLS, 2 Massachusetts Ave.,NE, Suite 3105, PSB, Washington, DC 20212, *jain_raj@bls.gov*

Key Words: Intervention Adjusted Series,, Intervention & Seasonally Adjusted Series, Trend, Smoothness

Using two basic assumptions of time-series analysis, a graphical test is proposed to assess if the quality of seasonal adjustment of a series is good or is not good. The test is applied to several seasonally adjusted series of the Consumer Price Index (CPI) using three methods of seasonal adjustment: X11, State Space Model Based Method and X13A (SEATS). The empirical results are presented in the form of comparative graphs.

Sutse Models And Multivariate Seasonal Adjustment

✦Filippo Moauro, EUROSTAT, 5, Rue Alphonse Weicker, Luxembourg, International L-2721 Luxembourg, *filippo.moauro@ec.europa.eu*

Key Words: SUTSE models, seasonal adjustment, state space methods, real time analysis

EUROSTAT has been always involved in the field of seasonal adjustment (SA): recent efforts went to the release of the ESS-Guidelines for harmonization of SA practices among European economic indicators, as well as the development of a new tool, DEMETRA+, based on the leading algorithms for SA, i.e. TRAMO&SEATS and X-12-ARIMA. This paper exploits an alternative methodology based on SUTSE models for multivariate SA in line with the ESS-Guidelines. It is discussed how to simultaneously perform SA estimates for a top aggregate of an economic indicator and their sub-components and how to assure consistency within the entire system of SA figures. Pre-treatment of time series is also treated in the study. The empirical application is devoted to the euro area industrial production index, presenting the results of a real time comparative study of the proposed approach with those produced by DEMETRA+.

Experiences With User-Defined Regressors For Seasonal Adjustment

✦Catherine Hood, Catherine Hood Consulting, 1090 Kennedy Creek Rd., Auburntown, TN 37016, *cath@catherinechhood.net*; Roxanne Feldpausch, Catherine Hood Consulting

Key Words: seasonal adjustment, moving holiday effects, trading day effects, X-12-ARIMA

The US Census Bureau's X-12-ARIMA and X-13-ARIMA-SEATS have built-in regressors for some US holidays and several types of trading day effects. However, there are many other effects that are not built-in. User-defined regressors are useful to estimate moving holidays not included in X-12-ARIMA, custom Easter effects, and custom trading day effects. Unfortunately, many users do not take the time to create their own regressors. The Census Bureau provides a program called "genhol" to help users generate their own holiday effects; however, this program can be daunting to users who prefer point-and-click software interfaces. We will demonstrate our user-friendly program that assists in defining regressors by including many built-in holidays, a flexible number of days before and after the holiday, and different kinds of one-variable trading day effects. Using time series from several different sectors, we demonstrate the value of seasonal adjustments that include custom holiday and trading day regressors. We evaluated the models using both modeling and seasonal adjustment diagnostics from X-12-ARIMA.

Unit Root Properties Of Seasonal Adjustment And Related Filters

✦William Robert Bell, U.S. Census Bureau, 4600 SIlver Hill Road, Room 5K142A, Washington, DC 20233, *William.R.Bell@census.gov*

Key Words: differencing, trend, polynomial, fixed seasonal effects, linear filter

Linear filters used in seasonal adjustment (model-based or from the X-11 method) contain unit root factors in the form of differencing operators and seasonal summation operators. The extent to which the various filters (seasonal, seasonal adjustment, trend, and irregular) contain these unit root factors determines whether the filters reproduce or annihilate (i) fixed seasonal effects, and (ii) polynomial functions of time. This paper catalogs which unit root factors are contained by the various filters for the most common approaches to model-based seasonal adjustment, and for X-11 seasonal adjustment with or without forecast extension. Both symmetric and asymmetric filters are considered.

Inference For Non-Stationary Time Series

✦Xiaoye Li, Penn State University, 126 E Nittany Ave. Apt. 4, State College, PA 16801 USA, *xul117@psu.edu*

Key Words: Change-point, Confidence interval, Strong invariance principle, Long-run variance, Non-stationary time series, Self-normalize

We study statistical inference for a class of non-stationary time series with time dependent variances. Based on a self-normalization technique, we address several inference problems, including self-normalized Central Limit Theorem, self-normalized cumulative sum test for change-point problem, long-run variance estimation through blockwise self-normalization, and self-normalization based wild bootstrap

for non-stationary time series. Monte Carlo simulation studies show that the proposed self-normalization based methods outperform stationarity based alternatives. We demonstrate the proposed methodology using two real data sets: annual mean precipitation rates in Seoul during 1771-2000, and quarterly U.S. Gross National Product growth rates during 1947-2002.

Quarterly Birth/Death Residual Forecasting

✦Nathan Clausen, US Bureau of Labor Statistics, DC 20212, *clausen.nathan@bls.gov*; Brian Dahlin, US Bureau of Labor Statistics

Key Words: Current Employment Statistics (CES), Birth/Death model, CES benchmark revision, Birth/Death residual forecast

Beginning in 2011, BLS began updating the Current Employment Statistics (CES) net birth/death model component of the estimation process more frequently, generating birth/death factors on a quarterly basis instead of annually. This allows CES to incorporate Quarterly Census of Employment and Wages (QCEW) data into the birth/death model as soon as it becomes available. This more frequent updating should help to reduce what is known as the "post-benchmark revision" in the CES series. Because the quarterly updating allows the most recent quarter of available QCEW data to be incorporated immediately, rather than at the end of the year, revisions between the initial birth/death residual forecasts and the revised birth/death forecasts are usually reduced. This paper documents research comparing both annual and quarterly methodologies to forecast net birth/death residuals using data from 2003 to 2009. The results show that the quarterly methodology would have led to smaller post-benchmark employment revisions for most years in the study.

247 Bayesian Monte Carlo Methodology

IMS, Section on Bayesian Statistical Science

Monday, August 1, 2:00 p.m.–3:50 p.m.

Mcmc Estimation Of Set Sizes On The Hypercube

✦Austen Wallace Head, Stanford University, 390 Serra Mall, Stanford, CA 94305, *ahead@stanford.edu*

Key Words: hypercube, sampling, MCMC, Metropolis, graph distributions, variance reduction

Markov chain methods are used to estimate the size of a set (a subset of a hypercube) which has a particular property (i.e. estimate μ=|S| for S={x:T(x)=1,x in O} where T:O?{0,1} is an indicator function and O is a hypercube). Estimates of μ are made by generating samples x in O with probability proportional to exp(T(x)fl) where fl is a parameter that we set to minimize the variance of m our estimate of μ. The estimate m based on r samples has a standard deviation approximately proportional to |O|/r. The Metropolis Hastings algorithm is used to generate samples, and the mixing time between samples to get near independence (based on total variation distance) is bounded as a function of fl and |O|. For small sets this method gives a substantially smaller variance than uniformly sampling on O. In addition to examples validating this methodology we show how this technique can be extended to problems which we do not already know how to answer. In graphs with n labelled vertices, this technique is used to estimate the distribution of the number of graphs with k triangles. Methodology is developed using this technique to estimate the how close an observed graph is to a threshold graph.

Spectral Analytic Comparisons For Data Augmentation

✦Vivekananda Roy, Iowa State University, Department of statistics, 3415 Snedecor Hall, Ames, IA 50010, *vroy@iastate.edu*

Key Words: data augmentation algorithm, spectrum, convergence rate, Markov chain, compact operator, eigenvalue

The data augmentation (DA) algorithm, though very useful in practice, often suffers from slow convergence. Hobert and Marchev (2008) recently introduced an alternative to DA algorithm, which we call sandwich DA (SDA) algorithm since it involves an extra move that is sandwiched between the two conditional draws of the DA algorithm. The SDA chain often converges much faster than the DA chain. In this paper we consider theoretical comparisons of DA and SDA algorithms. In particular, we prove that SDA is always as good as DA in terms of having smaller operator norm. If the Markov operator corresponding to the DA chain is compact and the extra move that is required in SDA is idempotent, which is often the case in practice, then the SDA is also compact and the spectrum of the SDA dominates that of the DA chain in the sense that all (ordered) eigenvalues of SDA are smaller than or equal to those corresponding eigenvalues of DA. We also present a necessary and sufficient condition that the extra move in SDA should satisfy for the operator norm of SDA to be strictly less than that of DA. We then consider some examples.

Extension Of Mixture Of Experts Model For Repeated Measures Data

✦Sungmin Myoung, Dept. of Biomedicl Informatics, Jungwon University, 5, Dongburi, Goesan-gun Chungcheongbuk-do, 367-805 Republic of Korea, *smmyoung@jwu.ac.kr*; Chungmo Nam, Dept. of Biostatistics, Yonsei University College of Medicine

Key Words: Mixture of Experts, Linear Mixed Effect model, Classification, EM-algorithm

The Mixture of experts (ME) is modular neural network architecture for supervised learning among a number of existed methods. This model can be considered as a mixture model that is consisted of the mixed distributions and weights in input variable x. Some of the researchers have been suggested to use of the multiple models for pattern classification and regression by ME method. However, the utilization of newly applied method is necessary in the repeated measures data. In this research, the mixture of experts is extended for repeated measures data. Moreover, cluster-specific effect is quantified via the linear mixed-effect model. To resolve this, firstly, we considered the construction of an expert, which has linear mixed-effect model, in ME. Afterward, the finding estimates were obtained for gating network and expert via EM-algorithm. The proposed model is more flexible than classical linear mixed-effect model in identifying several distinct clusters with different patterns. The simulated data and real data examples of kidney transplantation will be used to illustrate the feasibility of ME method.

Bayes Admissible Estimation Of Means In The Poisson Graphical Models

✦Hisayuki Hara, University of Tokyo, 7-3-1 Honfo Bunkyo-ku, Tokyo, 113-8656 Japan, *hara@tmi.t.u-tokyo.ac.jp*

Key Words: admissibility, Bayes, decomposable model, shrinkage estimation

We investigate the Bayes estimation of the means in Poisson decomposable graphical models. We give some classes of Bayes estimators improving on the maximum likelihood estimator under the normalized squared error loss. Both proper and improper priors are included in the proposed classes of priors. Concerning the generalized Bayes estimators with respect to the improper priors, we address their admissibility. Our estimator for the saturated model coincides with the estimator by Clevenson and Zidek (1975). In this sense our estimators are considered as a natural extension of the Clevenson-Zidek estimator to the graphical model.

A Bayesian Markov Switching Model For Change Point Detection In Sparse Causal Graphical Learning

✦Huijing Jiang, IBM Thomas J. Watson Research Center, , *huijiang@us.ibm.com*; Fei Liu, IBM Watson Research Center; Aurelie Lozano, IBM Thomas J. Watson Research Center

Key Words: Causal inference, Sparse graphical learning, Multivariate time series, Markov switching model, Group variable selection, Change point detection

Causal inference is an important topic in statistics and machine learning and has wide applicability ranging from biology to social sciences. Learning temporal graphical structures from multivariate time series data reveals important dependency relationship between current and past observations and is thus a key research focus for causal discovery. Most of the traditional methods assume a "static" temporal graph. Yet in many relevant applications, the underlying dependency structures may vary over time. In addition, with particular focus on the sparsity of the resulting causal graphical models, the lagged variables belonging to the same time series shall be included or excluded simultaneously. In this paper, we introduce a Markov switching vector autoregressive model to detect the structural changes of the causal relationship in multivariate time-series data. Our approach allows for such structural changes by a set of latent state variables modeled by a Markov process. At each state, we further impose the sparse structure of the causal graphical models through the hierarchical Bayesian group Lasso method. We demonstrate the value of our approach on simulated and real-world datasets

248 Model Selection & Forecasting ■●

Business and Economic Statistics Section

Monday, August 1, 2:00 p.m.–3:50 p.m.

Multi-Step Forecast Model Selection And Combination

✦Bruce E Hansen, University of Wisconsin, Department of Economics, 1180 Observatory Drive, Madison, WI 53706, *behansen@wisc.edu*

Key Words: MSFE, cross-validation, forecast combination, model averaging, information criterion, multi-step forecasting

This paper examines model selection and combination in the context of multi-step linear forecasting. We start by investigating multi-step mean squared forecast error (MSFE). We derive the bias of the in-sample sum of squared residuals as an estimator of the MSFE. We find that the bias is not generically a scale of the number of parameters, in contrast to the one-step-ahead forecasting case. Instead, the bias depends on the long-run variance of the forecast model. In consequence, standard information criterion (Akaike, FPE, Mallows and leave-one-out cross-validation) are biased estimators of the MSFE in multi-step forecast models. These criteria are generally under-penalizing for over-parameterization and this discrepancy is increasing in the forecast horizon. In contrast, we show that the leave-h-out cross validation criterion is an approximately unbiased estimator of the MSFE and is thus a suitable criterion for model selection. Leave-h-out is also suitable for selection of model weights for forecast combination. We provide strong simulation and empirical evidence in favor of weight selection by leave-h-out cross validation.

Subset Arma Model Selection Via Regularization

✦Kun Chen, University of Iowa, Department of Statistics and Actuarial Science, 241 Schaeffer Hall, The University of Iowa, Iowa City, IA 52242, *kun-chen@uiowa.edu*; Kung-Sik Chan, University of Iowa

Key Words: least squares regression, oracle properties, ridge regression, seasonal ARIMA models, sparsity, Lasso

Model selection is a critical aspect of subset autoregressive moving-average (ARMA) modelling. This is commonly done by subset selection methods, which may be computationally intensive and even impractical when the true ARMA orders of the underlying model are high. On the other hand, automatic variable selection methods based on regularization do not directly apply to this problem because the innovation process is latent. To solve this problem, we propose to identify the optimal subset ARMA model by fitting an adaptive Lasso regression of the time series on its lags and the lags of the residuals from a long autoregression fitted to the time-series data, where the residuals serve as proxies for the innovations. We show that, under some mild regularity conditions, the proposed method enjoys the oracle properties so that the method identifies the correct subset model with probability approaching one with increasing sample size, and that the estimators of the nonzero coefficients are asymptotically normal with the limiting distribution same as the case when the true zero coefficients are known a priori. We illustrate the new method with simulations and a real application.

Boosting As A Forecasting Method Using Large Datasets - Evidence For The Usa, The Euro Area And Germany

✦Teresa Buchen, Ifo Institute for Economic Research, Poschingerstr. 5, Munich, GA 81679 Germany, *buchen@ifo.de*; Klaus Wohlrabe, Ifo institute for Economic Research

Key Words: boosting, forecasting, large datasets

Recently, there has been rising interest in the use of large datasets for forecasting since both the availability of data and the computational power to handle them have strongly increased. Since traditional econometric methods are not suitable for incorporating a large number of predictors, new methods were developed. Boosting is a prediction method for high-dimensional data that was invented in the machine learning community. It is a stagewise additive modelling procedure, which, in a linear specication, becomes a variable selection device that sequentially adds the predictors with the largest contribution to the fit. The few related studies that exist in the macroeconometric literature are all restricted to U.S. data. By analysing large datasets for the USA, the euro area and Germany, this study evaluates the forecasting accuracy of boosting with respect to several key economic variables. We find that boosting mostly outperforms the benchmark, but the gains are largest in the medium run. Moreover, the forecast errors decrease when selecting a less noisy subset of the data. So the quality of data greatly determines the forecasting performance of boosting.

Obtaining Prediction Intervals For Farima Processes Using The Sieve Bootstrap

✦Maduka N Rupasinghe, Missouri University of Science and Technology, Department of Mathematics & Statistics, 400 West 12th Street, Rolla, MO 65409-0020, *mnrycd@mail.mst.edu*; Purna Mukhopadhyay, Univeristy of Kansas Medical Center; V A Samaranayake, Missouri university of Science and Technology

Key Words: Forecasting, Long Memory Processes, Fractionally Intergrated Processes, Model-based Bootstrap, ARFIMA processes

The Sieve Bootstrap method for constructing prediction intervals for invertible ARMA processes is based on re-samples of residuals obtained by fitting a finite degree autoregressive approximation to the process. The advantage of this approach is that it does not require the knowledge of the orders p and q associated with the ARMA model. The application of this method has been, up to now, limited to ARMA processes whose autoregressive polynomials do not have fractional roots. In this paper we adopt the sieve bootstrap method to obtain prediction intervals for ARFIMA (p, d, q) processes with 0< d< 0.5. The proposed procedure is a simpler alternative to an existing method, which requires the estimation of p, d, and q . Monte-Carlo simulation studies, carried out under the assumption of normal, t and exponential distributions for the innovations show near nominal coverage under most situations. Conservative coverages are obtained in cases a root of the non-fractional part of the autoregressive polynomial is close to unity.

Forecasting Time-To-Event Durations In Multi-Step Processes Using Non-Parametric Dynamic Survival Analysis

✦Nan Shao, IBM Research, , *nanshao@us.ibm.com*; Alejandro Veen, IBM Research

Key Words: time-to-event data, Kaplan-Meier estimate, time series, forecast

Accurate and timely forecasting of event occurrence plays an important role in numerous business applications such processing applications for loans or jobs. When the underlying distributions of the time-to-event durations change over time, the more recent data is more relevant for forecasting. However, a large proportion of recent events are still ongoing, and this causes censoring and thus limits the information. We propose a practical solution to the forecasting problem in such situations which includes a combination of non-parametric estimation and time series forecasting.

Bridge And Adaptive Regression Methods With Time Series Error

✦Taewook Lee, Hsnkuk University of Foreign Studies, , *twlee@hufs.ac.kr*

Key Words: Adaptive lasso, ARMA-GARCH models, Bridge regression, Consistency, Mixing error, Oracle property

Penalized regression methods have recently gained enormous attention in statistics and the field of machine learning due to their ability of addressing variable selection and multicollinearity. However, most of Penalized regression methods are limited to the case when the data are independently observed. In this paper, we consider bridge and adaptive lasso eestimators in linear regression models with mixing error and investigate their asymptotic properties such as consistency and the oracle property. In addition, we provide a computational algorithm that utilizes the local quadratic approximation and adaptively selects tuning parameters. As a special case of mixing error models, we consider linear regression models with ARMA-GARCH error models and introduce an improved algorithm that can not only produces small prediction errors but also accurately estimate an error model. We compare the performances of bridge estimators with other methods using simulated and real examples.

A New Efficiency Rate For Ols And Gls Estimators In Time Series Regressions

✦Jaechoul Lee, Boise State University, 1910 University De, Boise, ID 83725, *jaechlee@math.boisestate.edu*; Robert B. Lund, Clemson University

Key Words: Asymptotic variance, Autoregression, Convergence rate, Efficiency, Simple linear regression, Time series

When a straight line is fitted to data with autocorrelated errors, generalized least squares estimators of the trend slope and intercept are attractive as they are unbiased and of minimum variance. However, computing generalized least squares estimators is laborious as their form depends on the autocovariances of the regression errors. On the other hand, ordinary least squares estimators are easy to compute and do not involve error autocovariances. It has been known for 50 years that ordinary and generalized least squares estimators have the same asymptotic variance when the errors are second order stationary. Hence, little precision is gained by using generalized least squares estimators in stationary error settings. This paper revisits this classical issue, deriving explicit expressions for the generalized least squares estimators and their variances when the regression errors are an autoregressive process. These expressions are then used to show that ordinary least squares

methods are even more efficient than previously thought. Specifically, we show that the convergence rate of variance differences is one polynomial degree higher than previously explained.

249 Methods for Non Gaussian Data ■●

Section on Statistics and the Environment
Monday, August 1, 2:00 p.m.–3:50 p.m.

Robust Confidence Interval For Zero-Heavy Distribution

✦Mathew Anthony Cantos Rosales, COMSYS, 5220 Lovers Lane STE 200, Portage, MI 49002, *matewr@yahoo.com*; Magdalena Niewiadomska-Bugaj, Western Michigan University

Key Words: zero-heavy, lognormal distribution, Monte-Carlo, robust estimator, delta distribution, confidence interval

Zero-heavy data are very common in many disciplines like insurance, medical research, life sciences, marine sciences and engineering. In real life (e.g., in marine sciences), data that are said to be from lognormal distribution were often better fit by other skewed distributions (Myers and Pepin, 1990). This is in addition to the fact that goodness-of-fit test could not reliably detect departure from lognormality of the positive observation when sample size is small. In this paper, robustness of the interval estimators for the mean of lognormal distribution with a positive mass at zero was investigated and a new method was proposed. A comprehensive Monte-Carlo simulation study was performed and revealed that the proposed method outperforms other methods in terms of coverage probability and interval width when the data depart from the assumed model or are contaminated by extremely large values.

Exponentiated Sinh Cauchy Distribution With Applications

✦Kahadawala Cooray, Central Michigan University, Mt. Pleasant, MI 48859 USA, *coora1k@cmich.edu*

Key Words: Bimodal and unimodal data, Cauchy distribution, Hyperbolic secant distribution

The exponentiated sinh Cauchy distribution is characterized by four parameters: location, scale, symmetry, and asymmetry. The symmetric parameter preserves the symmetry of the distribution by producing both bimodal and unimodal densities having coefficient of kurtosis values range from one to positive infinity. The asymmetric parameter changes the symmetry of the distribution by producing both positively and negatively skewed densities having coefficient of skewness values range from negative infinity to positive infinity. Bimodality, skewness, and kurtosis properties of this regular distribution are presented. In addition, relations to some well-known distributions are examined in terms of skewness and kurtosis by constructing aliases of the proposed distribution on the symmetric and asymmetric parameter plane. The maximum likelihood parameter estimation technique is discussed, and examples are provided, analyzed, and compared based on data from environmental sciences to illustrate the flexibility of the distribution for modeling bimodal and unimodal data.

Lognormal Block Kriging For Multivariate Spatial Processes

✦Mona Abdullah Alduailij, Western Michigan University, 4435 Wimbleton way, Kalamazoo, MI 49009, *malduailej@yahoo.com*; Rajib Paul, Western Michigan University

Key Words: Cross-covariance, Dimension reduction, Nonstationarity, Remote-Sensing

Remote-Sensing datasets obtained from satellite are very large and often require dimension reduction techniques for statistical analysis. We analyze datasets on three cloud properties. (i) Cloud Optical Thickness, (ii) Cloud Top Pressure, and (iii) Cloud Water Path, obtained from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS). These datasets exhibit skewness and log-transformation provides numerical stability. We derive the optimal predictor for the unobserved locations after taking into consideration the nonstationarity, spatial and cross-covariances of these multivariate spatial processes. We also produce maps for the root-mean squared prediction errors to assess the goodness of our predictions.

Gaussian Subordination Models On A Lattice With Environmental Applications

✦Sucharita Ghosh, Swiss Federal Research Institute WSL, Zuercherstrasse 111, Birmensdorf, International CH-8903 Switzerland, *rita.ghosh@wsl.ch*

Key Words: Spatial data, Smoothing, Long-range dependence, Large deviation, Forestry, Ecology

Suppose that spatial observations Y(i,j) occur on a lattice (i,j), i= 1,...,n, j=1,2,.,m such that the data are Gaussian subordinated via a function G that is unknown and arbitrary except that it allows for a Hermite polynomial expansion. The advantage of this model is that it allows for non-Gaussianity of the data and that the shape of the underlying probability distribution may be location dependent. We consider various correlation types and in particular short memory and long-memory correlations and address two topics: (a) the nonparametric regression problem where the errors are Gaussian subordinated as described above and (b) the species count problem where the background process that is decisive of species occurrence is a Gaussian subordinated process. Generalization to the case when the data are irregularly spaced in space are also considered. We discuss asymptotic results and some real data applications from environmental monitoring.

Bayesian Probit Regression For Multicategory Spatial Data

✦Candace Berrett, Brigham Young University, , *cberrett@stat.byu.edu*; Catherine A. Calder, The Ohio State University

Key Words: spatial statistics, categorical data, latent variable methods, data augmentation

Albert and Chib (1993)'s latent variable representation of the Bayesian probit regression model for categorical outcomes is widely recognized to facilitate model fitting. This representation has also been used in various settings to incorporate residual dependence into regression models with discrete outcomes. In this talk, we further extend this latent variable strategy to specify models for multicategory spatially-dependent outcomes. In particular, we discuss parameter identifiability

issues in the latent mean specification and introduce covariance structures for describing the cross spatial/category residual dependence. We also consider data augmentation MCMC strategies for improving the efficiency of model fitting algorithms. Finally, we illustrate the proposed modeling framework through an analysis of land-cover/land-use observations taken over mainland Southeast Asia.

Using Aic On Transformed Data

✦Mark C. Otto, U.S. Fish and Wildlife Service, 2302 Lackawanna St, Adelphi,, 20783, *Mark_otto@fws.gov*

Key Words: model fit, model selection

Akaike's Information Criterion (AIC) uses information theory to compare statistical models for the same data. AIC can be used to compare model s with that have different distributions or with regression variables that are not nested. AIC can also be used to compare models with different transformations. David Findley applied a transformation of variables to derive an AIC on the original scale for models on transformed data. The AICs differ only in their Jacobian of transformation. This calculation has been used on time series models in the Census Bureau's RegARIMA, time series and regression modeling program and is presented in G. Kitagawa's Introduction to Time Series Modeling but is not otherwise widely used. AIC for transformations can be used on regression and generalized linear models. We show its effectiveness on simulated and real data.

Principal Component Analysis For Interval Data - Perspective

✦Lynne Billard, University of Georgia, Department of Statistics, University of Georgia, Athens, GA 30602 USA, *lynne@stat.uga.edu*; Jennifer Le-Rademacher, Medical College of Wisconsin

Key Words: vertices method, symbolic method, vertex contribution

Although interval data occur naturally in their own right (such as species data), they will become more and more ubiquitous as contemporary computer capabilities generate massively large data sets necessitating aggregation in some way. We look at this phenomena first. Then we provide a perspective on available methods for principal component analyses (PCA) of interval data, and how results differ from and expand upon those for traditional PCA on classical point observations.

250 Survey Sampling Methodology in Epidemiology

Section on Statistics in Epidemiology, Section on Health Policy Statistics, Section on Survey Research Methods, Statistics Without Borders, Scientific and Public Affairs Advisory Committee

Monday, August 1, 2:00 p.m.–3:50 p.m.

Relationship Of Smoking And Cancer In Women: A National Health And Nutrition Examination Survey (Nhanes) 2005-2006 Study

✦Sunita Ghosh, Alberta Health Services-Cancer Care, 11560 University Ave., Edmonton, Alberta, T6G 1Z2 Canada, *sunita.ghosh@ualberta.ca*; Qiahao Zhu, Alberta Health Services-Cancer Care

Key Words: NHANES, cancer, smoking, GEE, alcohol

The relationship of cancer and smoking is complex and not very clear. Further more, there have not been many studies exploring the relationship of smoking and cancer. In the present paper we investigate the relationship of cancer (breast, ovarian, uterus and cervix cancer) and smoking history of patients using the National Health and Nutrition Examination Survey (NHANES) data. Apart from smoking history we also studied other important factors known to be associated with cancer: body mass index, age, alcohol consumption and exposure to second hand smoke. Since NHANES data set uses a complex survey design, hence we used generalized estimating equation method to account for the stratification, clustering and unequal weighting. The results of the analysis suggest that smoking is positively associated with risk of cancer. The odds of having cancer are much higher in women who currently smoke or are ex-heavy smokers compared to non-smokers. These results were adjusted for alcohol consumption and age. The results of the present study confirm our hypothesis that cigarette smoking is positively associated with cancer among women.

Estimating Family Aggregation Of Disease In Cross-Sectional Surveys

✦Barry Ira Graubard, National Cancer Institute, Biostatistics Branch, 6120 Executive Blvd, Rm 8024, Bethesda, MD 20852, *graubarb@mail.nih.gov*; Monroe G Sirken, National Center for Health Statistics

Key Words: genetic studies, disease aggregation, sample surveys, network sampling

Family-based genetic studies often use convenience samples of families that are constructed from probands with diseases of interest. Using these samples, measures of familial aggregation of these diseases for certain relationships (e.g., siblings) are compared to familial aggregation of probands without disease or to the general population to determine if there may be a genetic cause. Two popular measures of aggregation are the Reccurence Risk Ratio and Intraclass Correlation Coefficients. Since family-based studies are not random samples of families, estimates of aggregation may be biased. Probability samples of individuals obtained in surveys such as the National Health Interview Survey (NHIS) can collect disease statuses of relatives of surveyed individuals. Proper weighting that accounts for probabilities of reporting family members of surveyed individuals provides unbiased estimates of aggregation. This paper illustrates the estimation of family aggregation of diabetes in the NHIS using sibling relationships. These aggregation measures have potential utility in analyzing familial distributions of conditions that could be associated with social and ecological risks.

Mantel-Haenszel Estimators for Complex Survey Data

✦Babette Brumback, University of Florida, Department of Biostatistics, Gainesville, FL 32610, *brumback@phhp.ufl.edu*; Zhulin He, University of Florida

Key Words: odds ratio, complex sampling designs, conditional logistic regression, sparse-data limiting models, cluster sampling

The SUDAAN CROSSTAB procedure provides an adaptation of the Mantel-Haenszel estimator of a common odds ratio for complex survey data. The estimator is consistent for large-strata limiting models, in which the number of strata remains fixed and the sample sizes within strata increase to infinity. However, we show via simulation and counterexamples that the estimator is inconsistent for sparse-data limiting models, in which the number of strata (more appropriately termed as clusters) increases to infinity but the sample sizes within cluster remain fixed. We also propose an alternative estimator that is consistent for sparse-data limiting models satisfying a positivity condition, but not for large-strata limiting models. We compare the estimators with each other and with recent adaptations of conditional logistic regression for complex survey data.

Analysis Of Case-Control Studies With Sample Weights

✦Victoria Landsman, National Cancer Institute, Bethesda, MD 20852, *landsmanv@mail.nih.gov*; Barry Ira Graubard, National Cancer Institute

Key Words: Informative sampling,, Modeling sample weights, Weighted estimating equations, Sample distribution

Analysis of population-based case-control studies with complex sampling designs is challenging because the sample selection probabilities (and, therefore, the sample weights) depend on the response variable and covariates. In this paper we propose a new semi-parametric weighted estimator which incorporates modeling of the sample expectations of the weights into a design-based framework. The estimator has a theoretical basis and is robust to model misspecification. We describe also the sample pseudo maximum likelihood estimator for inference from case-control data. This approach generalizes the Breslow and Cain (1988) estimator and it can be applied to the data from complex sampling designs, including cluster samples. We discuss benefits and limitations of each of the two proposed estimators emphasizing efficiency and robustness. We compare the final sample properties of the two new estimators and the existing weighted estimators in simulations under different sampling plans. We apply the methods to the National Cancer Institute's U.S. Kidney Cancer Case-Control Study in order to identify risk factors for kidney cancer.

Integrating Data From Two Surveys To Estimate The Prevalence Of Cervical Cancer Screening And Its Associated Factors In The U.S.-Mexico Border Region

✦Ruben Smith, Centers for Disease Control and Prevention, 4770 Buford Highway, NE, Mailstop K-22, Atlanta, GA 30341, *eyb4@cdc.gov*; Dyanne Herrera, Centers for Disease Control and Prevention; Emily Schiefelbein, Texas Department of State Health Services; Jill McDonald, Centers for Disease Control and Prevention ; Gita Mirchandani, Texas Department of State Health Services

Key Words: Complex survey data, data pooling, subpopulation analysis, logistic regression

A binational surveillance system for the U.S.-Mexico border does not exist and reliable estimates for cervical cancer screening are not available for the region. We used comparable 2006 data from the BRFSS, a state based telephone survey, and the ENSANut, a state representative area probability sample and face-to-face survey, to estimate the prevalence of cervical cancer screening, and to describe associated factors in women aged 20-70 years without previous hysterectomy living in the border. The two surveys are independent and the target subpopulation for this study is contained in the combination of the populations targeted for the surveys. We considered each survey target population as a super stratum and then the data from the surveys were pooled together and domain type analyses were performed using SUDAAN. Among women aged 20-70 years without previous hysterectomy, 47% (95%CI=44.5-50.2) reported having a cervical cancer screening within the last year. Health insurance (AOR=1.9, 95%CI=1.5-2.4), marriage (AOR=1.8; 95%CI=1.4-2.3) and living on the US side of the border (AOR=3.5, 95%CI=2.8-4.4) were positively associated with having a cervical cancer screening.

Survey Design And Results From Unicef Project In Sierra Leone

✦Theresa Diaz, UNICEF, , *tdiaz@unicef.org*; Sowmya R Rao, UMass Medical School; John Wolkon, CDC; Gary Shapiro, Statistics Without Border; Peter S. Bangura, Statistics Sierra Leone; John Baimba, Statistics Sierra Leone

Key Words: childhood mortality, surveys, cluster sample

In Sierra Leone, UNICEF is implementing and evaluating Community Health Volunteers strategy to treat Malaria, Diarrhea, and Pneumonia in children< 5 years of age. As part of this evaluation, a baseline survey was conducted in 2010. A follow-up using the same sampling design and methodology is planned for 2012. This paper describes the sample design and preliminary analysis results. The design is a household cluster survey in two intervention and two control districts that were chosen based on pre-determined criteria. A cluster sample of 3000 households was then selected from each district using a probability proportional to size sampling scheme. Population-based interviews were conducted in person. Both direct and indirect mortality rates were computed. The results indicate that the intervention and control districts did not differ significantly in socio-economic factors, disease prevalence or health seeking behaviors. Mortality rates were lower than in previous surveys indicating declining rates of childhood mortality in Sierra Leone.

Use Of Gps-Enabled Mobile Devices To Conduct Surveys

✦Sowmya R Rao, UMass Medical School, 55 Lake Avenue North, Worcester, MA 01655, *sowmya.rao@umassmed.edu*; Gary Shapiro, Statistics Without Border; Theresa Diaz, UNICEF

Key Words: Surveys, GPS-enabled mobile device, PDA

Sampling households (HHs) based on geographic location for surveys in countries where such information is unavailable can be challenging. The use of Geographic Information Systems can help overcome this challenge. As part of a UNICEF project in Sierra Leone, to implement and evaluate Community Health Volunteers strategy to treat Malaria, Diarrhea, and Pneumonia in children under five years of age, a baseline survey of 6000 HHs was conducted. Global Positioning System (GPS) enabled mobile devices were used to enumerate HHs using GPS sample. The HHs were then selected using a simple random sampling scheme which was coded into the mobile device (e.g., PDA). The skip patterns for the survey were also programmed into the PDA, vastly improving quality control of the data. These data were then uploaded

into an ACCESS database, making them immediately available for analysis. Another advantage of using PDAs was to examine whether the interviews were actually conducted at the selected household, by comparing the GPS recorded location of the household at enumeration and at the time of the interview. The survey was highly successful with very high response rates and good quality data.

251 Topic Contributed Oral Poster Presentations: Section on Statistical Graphics

Section on Statistical Graphics, Section on Statistical Computing
Monday, August 1, 2:00 p.m.–3:50 p.m.

Oil At Sea

✦Antony Unwin, Uni Augsburg, Augsburg, 86482 Germany, *unwin@math.uni-augsburg.de*

Entrant for the Data Expo 2011

The Effect Of The Oil Spill On Nature

Patricia Millan, California State University, East Bay; ✦Luis Fernando Campos, University of California, Berkeley, 2914 A Deakin St, Berkeley, CA 95705, *lfcampos@berkeley.edu*

Key Words: Oil Spill, gulf, graphics

The BP oil spill, which arose from an explosion at the Deepwater Horizon rig, has left many unanswered questions related to the impact on the animals that inhabit the gulf region. The data collected by the Environmental Protection Agency (EPA), US Fish and Wildlife Service, and the National Ocean and Atmospheric Administration (NOAA) has allowed us to analyze and provide graphical summaries to help answer some of these following questions: What species of birds were the most affected and where were they when found? Specifically, is there a trend in the species of birds that were reported "not visibly oiled" and "visibly oiled"? Is this trend visible based on where they were found, or whether the birds were reported dead or alive? We will also explore how the species are affected as the oil spill progressed. The graphical summaries provided will help the general audience truly comprehend the impact the oil spill had on the animal inhabitants of the gulf region.

Visualizing Deepwater Horizon Oil Spill Data With R

✦Andras Low, ELTE TaTK, Budapest, International Hungary, *low.andras@gmail.com*

Key Words: visualizing, spatiotemporal

NOAA and EPA have a tremendous amount of data on temperature, salinity and petrochemical products but they are also very sparse, spatially scattered. These data can be exploited for many reasons, e.g., the extent of the oil spill visible in measurements on temperature and salinity; the consistency of temperature and salinity measurements between measuring devices; pattern in the temperature and salinity measurements that might indicate presence of oil. In this poster, I will describe how to use a combination of tools in R to seamlessly analyze and visualize these spatiotemporal effects.

Capturing The Deepwater Horizon Oil Spill Effects Through The Use Of Interactive Graphics

✦Nicolle Clements, Temple University, , *tuc37728@temple.edu*

Key Words: BP oil spill, 3D plots, oil, Deepwater Horizon, rotatable plots, heat maps

Capturing the Deepwater Horizon Oil Spill Effects Through the Use of Interactive Graphics Nicolle Clements, Temple University The Deepwater Horizon explosion on April 20, 2010 has numerous impacts that continue after the well was capped. With an estimated that 53,000 barrels per day escaping over a three month time span, graphical models are a great resource to answer the questions such as: Where did the oil go? Has it impacted the surrounding wildlife? What are the water chemical consequences? To answer some of these questions, a variety of interactive graphical models will be explored. Specifically, 3-D rotatable plots, heat maps, and surface plots can reveal data measurements at various longitude, latitude, and ocean depth.

Health Effects On Clean-Up Workers Of Deepwater Horizon

✦Cathy Furlong, StatAid, 6930 Carroll Ave. Ste. 420, Takoma Park, MD 20912, *cathy@stataid.org*; Juan Carlos Rosa, StatAid; Chelsea Wald, StatAid; Mary McGraw-Gross, StatAid

Key Words: Health, Deepwater Horizon, Human Rights, Directional changes in oil spill, Statistical graphics

Using a human-rights framework to evaluate the Deepwater Horizon oil spill, the StatAid team will graph changes in health data from on- and off-shore clean-up teams in relation to changes in EPA data and directional changes of the oil spill. Additional data was downloaded from http://www.nodc.noaa.gov/OC5/SELECT/dbsearch/dbsearch.html and http://www.cdc.gov/niosh/topics/oilspillresponse/.

Where Does The Oil Go?

✦Tianxi Li, Stanford University, 119 Quillen Ct, Apt 214, Stanford, CA 94305, *tianxilicb@gmail.com*; Gao Chao, Yale University; Meng Xu, Nankai University, Department of Environmental Science

Key Words: boosted trees, regression, classification

Our analysis can be divided into two parts. On one hand, we combined the floats and gliders data together, and tried to see the distribution of salinity and temperature. We also incorporate the birds and turtles information into the data, and try to see the impact of different factors on the animal survival situations over time. The main way to work on all of these predictors together is boosted regression trees and classification trees, as well as GAM models. To model the inherent correlations between predictors well, we also make certain transformation by regressions as well. Then by such methods, we showed how the eco-system in the area varies with time. This can be interpreted as the impact of oil, which leads to a roughly estimation of oil's movement. On the other hand, we also delved into the birds and turtles data set as a separate problem, which recorded the oil conditions and live conditions on different locations and dates. Classification boundaries for live conditions and oil conditions were plotted by different algorithms, and shows how the oil moved during the period of data.

A Dynamic Display Of The Bp Oil Spill Using Google Maps And Animation Techniques

✦Michelle S. Marcovitz, Temple University, 390K Speakman Hall, 1810 N. 13th Street, Philadelphia, 19122, *tua03619@temple.edu*; Bu Hyoung Lee, Temple University

This poster involves tracking the path of the BP oil spill with use of Google Maps. The goal is to show a relationship between the spill and the presence of harmful chemicals in the Gulf of Mexico and to chart the extent of the spill over time. This will be accomplished by the smoothing method to capture important patterns. After smoothing the data, animated graphs overlaid on maps of the gulf region display the virtual movement of the spill over time and trends between the presence of oil and harmful chemicals. This research will provide visible and dynamic images to explain how the spill has contaminated the gulf region with the smoothing method and animation techniques.

The Days After Deepwater Horizon Spill - A Graphical Summary Of A Catastrophe

✦Walter Hickey, College of William and Mary, CSU 4378, PO Box 8795, Williamsburg, VA , *wlhickey@email.wm.edu*; Bimal Parakkal, College of William and Mary

The data available through NOAA, the EPA, and the US Fish and Wildlife Service regarding the Deepwater Horizon oil spill of 2010 covers a broad spread, examining impact on wildlife, fishery contamination, and temperature and salinity results for a range of depths and regions. By identifying areas where there is a dissonance between temperature and salinity percentiles we may identify regions of interest. By analyzing these regions of interest over several datasets, we may develop a conclusion regarding the role that temperature and salinity play in identifying oil contamination. By comparing current fishery contamination with baseline toxicity we identify the impact of the spill on organisms and the extent of contamination. Likewise, by analyzing the rate and conditions of avian and reptilian morbidity and mortality across the gulf we find additional answers regarding the extent and impact of the spill.

Visualizing The Data Age From Mobile In-Situ Sensors

✦Daniel N¸st, 52North GmbH, Martin-Luther-King-Weg 24, 48155, Muenster, Germany, *daniel.nuest@uni-muenster.de*; Edzer Pebesma, Institute for Geoinformatics, University of Muenster

Key Words: interpolation, sample planning, oil spill, moving sensors

Mobile in-situ sensors observe phenomena at different times and different locations. If the phenomenon is dynamic, for instance because it moves or its characteristics change over time, all inferences that are drawn from the observation must be critically reviewed with regard to the respective age of the measurements. This is even more the case when a phenomenon is being monitored in near real-time. We think that this problem applies to a variety of measurements taken around the oil spill, e.g. temperature and salinity by boats and gliders. Our approach comprises a visualisation of both interpolated values and observation points in a dynamic three-dimensional setting. The sparse nature of the measurements makes calculations vulnerable to misinterpretation if not combined with information on temporal and spatial aspects of the measured events. We see this kind of metadata, provided alongside the data visualization, as crucial to understanding an observed phenomenon. This might not directly answer the question whether there is evidence for a contamination, or where the oil went, but where one can actually make such investigations as the data could be sufficient, or where addition

Data Expo 2011: The Deepwater Horizon Disaster - An Oily Affair

✦Anvar A. Suyundikov, Utah State University, Department of Mathematics and Statistics, 3900 Old Main Hill, Logan, UT 84322-3900, *anvar_suyundikov@yahoo.com*; Colby W. Brungard, Utah State University; Lynsie M. Daley, Utah State University; Placede Judicae Gangnang Fosso, Utah State University; Huong La, Utah State University ; Chang Li, Utah State University; Yuanzhi Li, Utah State University; Nathan D. Voge, Utah State University; Juergen Symanzik, Utah State University

Key Words: Exploratory Data Analysis, Graphical Analysis, Oil Spill

The explosion on the Deepwater Horizon oil drilling rig on April 20, 2010, that killed eleven workers and injured another 17, is the basis of one of the biggest environmental disasters in the Gulf of Mexico. In this poster, we take a graphical look at the spatial and temporal development of the resulting oil spill and its environmental effects on local economy and wildlife.

Graphical Investigation Of The Oil Spill

✦Heike Hofmann, Iowa State University, , *hofmann@iastate.edu*

Key Words: Data Expo, Oil Spill Data, Graphics

We will present results from our investigation of the oil spill data in graphical from. This contribution is grown from the Stat 480 course project at Iowa State University.

Deepwater Horizon Oil In The Gulf

✦Lou Bajuk, TIBCO Software Inc., 1700 Westlake N., #500, Seattle, WA 98109, *lbajuk@tibco.com*; Stephen Kaluzny, TIBCO Software Inc

Key Words: data expo 2011, gulf oil, Spotfire, R

What happened to the 5 million barrels of crude oil that flowed into the Gulf of Mexico from the Deepwater Horizon oil spill? We will examine the spread of the oil and its effects on the Gulf using a high interaction graphics system (Spotfire) tightly coupled to a statistics engine (R). The system allows us to link different views of the data including spatial maps. The ability to fit and display simple models with the system and to drill down to specific geographic areas and subsets of the date add to our understanding of the effects of the oil.

Effects Of Oil Spill On Birds

✦Aida Yazdanparast, Cal State East Bay, 1901 Harder Road,, Internatiol house#641D,mailbox#1164, Hayward, CA 94542, *ayazdanparast@horizon.csueastbay.edu*; Tony Tran, Cal State East Bay

Key Words: Bird, Oil spill, Graphical summary

The Deepwater Horizon oil spill was an ecologically devastating event in the Gulf of Mexico, which saw an estimated release of over 4 million barrels of oil after flowing for three months in 2010. The impact

of the spill persisted despite the well's capping. Understanding the progression and making use of the details in this event is the first step in planning to improve response time and reaction strategy for future disasters relating to a local avian population. The aim of this project is to dynamically illustrate the important features of the data set utilizing a blend of analytics and graphics executed through R and Tableau software. In our data set, we focus on 7,229 birds that were documented between May and October. We will explore predicted probabilities as well as drill into the layers on a variety of canvases to uncover relevant information unseen through raw data for policymakers, environmental planners and ecology researchers, and others fascinated by the importance of birds in our ecosystem.

252 Contributed Oral Poster Presentations: Section on Health Policy Statistics

Section on Health Policy Statistics
Monday, August 1, 2:00 p.m.–3:50 p.m.

A Reexamination Of Nigeria'S Health Insurance Policy

✦Godson Mesike, University of Lagos, Department of Actuarial Science and Insurance, Faculty of Business Adminstration, Lagos, International N1014 Nigeria, *mesikegodson@yahoo.co.uk*; Ismaila Adeleke, University of Lagos; Ade Ibiwoye, University of Lagos; Dallah Hamadu, University of Lagos

Key Words: National Health Insurance Scheme, private sector participation, health care cost, fee structure,

The National Health Insurance Scheme aims to bring about a comprehensive health care to every Nigerian The main thrust is to protect families from the financial hardship of huge medical bills, ensure equitable distribution of health care costs among different income groups, maintain high standard of health care services delivery within the scheme and harness private sector participation in the provision of health care services. However, ten years on the objectives seem to be elusive. Many of the citizenry are jaded by the failure of earlier programs like the social insurance trust fund (NSITF) and the housing fund (NHF). However, health care is paramount to well being and national productivity and can not therefore be handled lightly. This makes it imperative to examine the existing program for cost effectiveness. This study appraises the funding of the NHIS scheme to determine if the present fee structure can sustain the program and makes policy recommendation.

Evaluation Of Consumer Panel Survey Data For Public Health Communication Planning: An Analysis Of Sixteen Years Of Annual Survey Data From 1995 - 2010

✦William E. Pollard, Centers for Disease Control and Prevention (CDC), 2554 Circlewood Rd., NE, Atlanta, GA 30345, *bdp4@cdc.gov*

Key Words: Health Survey Methods, Public Health Communication

Consumer survey data on media habits, lifestyle, and product use can provide useful information for understanding audiences and developing messages for health communication planning. A widely-used methodology for conducting consumer surveys involves sampling from panels of individuals and households that have been prerecruited to participate in surveys and other market research. Commercial research firms develop and maintain panels of several hundred thousand or more individuals across a full range of demographics for use in market research. In this presentation, data from an annual consumer health survey from 1995 to 2010 with 3,000 to 4,000 respondents per year are examined. Comparisons are made with other national survey data from these same years obtained through national probability sampling methods. Overall response levels, demographic breakdowns, and trends over time are examined. The literature on the use of panels is referenced and the implications for heath communication planning are discussed.

253 Contributed Oral Poster Presentations: Section on Physical and Engineering Sciences

Section on Physical and Engineering Sciences
Monday, August 1, 2:00 p.m.–3:50 p.m.

A Construction Method of Two-Level Split-Lot Designs Based on Nonregular Fractional Factorial Designs

✦Tena Katsaounis, Ohio State University, Ovalwood 378, 1680 University Drive, Mansfield, OH 44906 USA, *katsaounis.1@osu.edu*

Key Words: experimental design, factorial design, split-plot design, multi-stage design, nonregular design, optimal design

A construction method, based on nonregular fractions, is presented for small two-level designs which are performed in more than one stage andwhich have a split-plot structure in each stage. These designs are called "non-regular split-lot designs". The construction method yields designs with a small number of aliased factorial effects and a small number of factorial effects that are confounded to whole plot contrasts. A key concept is that of "relatively independent subspaces",which allows construction of (ideally disjoint) subspaces whose elements are associated with the factorial effects in the different stages of a design.

Evacuation of a Smoky Room

✦Guillermo Frank, University of Buenos Aires, PabellÛn I, Ciudad Universitaria, Buenos Aires, 1428 Argentina, *accesofrank@yahoo.com*; Claudio Dorso, University of Buenos Aires

Key Words: panic evacuation, smoky room

The escaping process of a panicking crowd involves very complex phenomena. The observed slowing down in the evacuation is actually known to be a consequence of the "faster is slower" effect. This effect states that as pedestrians desire to reach the exit increases, the clogging phenomena delays the time to get out of the room. A particularly harmful situation occurs when the environment is so smoky that pedestrians can not find the way out. They move like a "herd", towards the mean direction of the nearby pedestrians. The statistical behaviour

of the crowd shows a new kind of slowing down. Our investigation focuses on this source of inefficiency. We further compare this situation with the "faster is slower" effect.

Resampling In Generalized Linear Models

✦Maher Qumsiyeh, University of Dayton, Department of Mathematics, Dayton, OH 45469-2316, *maher.qumsiyeh@notes.udayton.edu*

Key Words: Experimental Design, Discrete Responses, Transformations, Generalized Linear Model

Frequently in experimental situations observations that are not normally distributed occur. A common situation is when the responses are discrete in nature such as counts. One way to analyze such experiments is to use a transformation for the responses. Another is to use a link function based on a generalized linear model approach. In this paper, we will use re-sampling as an alternative method to analyze such data. We will compare our results with those provided by the previous two methods.

Application Of Clustering In Building Mass Spectral Libraries

✦Xiaoyu Yang, National Institute of Standards and Technology, 100 Bureau Drive, Stop 8320, Gaithersburg, MD 20899, *xiaoyu.yang@nist.gov*; Pedatsur Neta, National Institute of Standards and Technology; Yamil SimÛn-Manso, National Institute of Standards and Technology; Stephen Stein, National Institute of Standards and Technology

Key Words: distance-based clustering, spectrum similarity, adjusted dot product, mass spectrum, mass spectral library, proteomic and metabolomic data

Tandem mass spectrometry (MS/MS) is becoming a major technique in proteomic and metabolomic studies. Spectrum library searching compares each experimental spectrum with reference spectra in the library. It has been proposed as an effective method for identifying small molecules in metabolomics and peptide identification in proteomics. Combining multiple similar spectra to create high quality consensus spectra is a key step in building mass spectral libraries. We developed a distance-based clustering algorithm that can efficiently create consensus spectra from MS/MS spectra. Dot product was used to evaluate spectrum similarities and further improved for more accurate clustering by penalizing its value based on peak intensities and intensity ratios. Clusters were also refined by reducing peak merging and peak splitting. This method has been validated with more than 10,000 spectra of more than 1000 metabolites and 3000 peptides. It has been applied to building NIST MS/MS peptide and metabolite libraries.

Deconvolution And Estimation Of Multiple Planar Signals

✦Joshua Kerr, CSU East Bay, Dept. of Statistics & Biostatistics, 25800 Carlos Bee Blvd., Hayward, CA 94542, *joshua.kerr@csueastbay.edu*

Key Words: simulation, seismic array, multiple signals, deconvolution

The task of extricating multiple signals arriving in planar fashion to noise-polluted sensors at an array site is an arduous one. This poster illustrates a method to calculate the number of signals present, where they are coming from and how fast they are moving. Asymptotic results on estimators are investigated to achieve interval estimates on signal attributes. The adequacy of the method will be judged through an application with real teleseismic event data accompanied with simulations.

Lifetime Predictive Density Estimation In Accelerated Degradation Testing For Lognormal Response Distributions With An Arrhenius Rate Relationship

✦Steven Michael Alferink, Missouri University of Science and Technology, 1612 Yarmouth Lane, Mansfield, TX 76063, *steve@mst.edu*; V A Samaranayake, Missouri university of Science and Technology

Key Words: Maximum Likelihood Predictive Density, Accelerated Degradation Testing, Lognormal Distribution, Prediction Bounds, Bootstrap Intervals

A relatively simple method is proposed for obtaining a predictive density for the lifetime of a future specimen at the design stress level in an accelerated degradation model. The model assumes the natural logarithm of a response variable has a normal distribution with a mean that follows the Arrhenius relationship and a standard deviation that is dependent on the accelerating stress, yet independent of time. The model uses the accelerated degradation of specimens at two or more accelerating stress levels where each specimen is measured only once. Failure is assumed to occur when the response variable crosses a predefined threshold. This method is based on the Maximum Likelihood Predictive Density approach first proposed by Lejeune and Faulkenberry. The use of the percentiles of the predictive density as prediction bounds for the lifetime of a single future specimen is examined using Monte Carlo simulation. Comparisons are also made with the prediction bounds obtained using the traditional Maximum Likelihood Estimator and bootstrap techniques.

Step-Stress Accelerated Degradation Testing For Reliability Of Solar Reflector

✦Jinsuk Lee, National Renewable Energy Laboratory, 1617 Cole Blvd, MS 1608, Golden, CO 80401, *jinsuk.lee@nrel.gov*; Ryan Elmore, National Renewable Energy Laboratory; Cheryl Kennedy, National Renewable Energy Laboratory; Wesley Jones, National Renewable Energy Laboratory

Key Words: Accelerated Testing, CSP, step-stress

Concentrating solar power (CSP) technologies use large mirrors to concentrate sunlight and convert the thermal energy collected to electricity. Commercialization of CSP technologies requires advanced optical materials that are low in cost and maintain high optical performance for lifetimes of 30 years under severe outdoor environments. In order to meet these stringent requirements, step-stress accelerated degradation testing (SSADT) technique was introduced to the CSP program at the National Renewable Energy Lab (NREL). In this talk, we will discuss the underlying statistical and physical theory used to construct the SSADT methodology for the CSP program and how it can be used to predict service lifetime, including warranty time and mean-time-to-

failure, failure/degradation rates at use-conditions, activation energy, acceleration factors, and upper limit level of stress. We will also discuss the accelerated environmental conditions that are encountered in typical accelerated CSP testing programs.

Analyzing The Effects Of L.E.D. Traffic Signals On Urban Intersection Safety

✦Peter W Hovey, University of Dayton, 300 College Park Dr, Dayton, OH 45469-2316, *Peter.Hovey@notes.udayton.edu*; Deogratias Eustace, University of Dayton; Deogratias Eustace, University of Dayton

Key Words: Traffic, Safety, Empirical Bayes, LED

The use of light emitting diodes (LED's) in traffic signals has become widespread over the past decade. Energy efficiency and long service life are the often-cited reasons for converting from incandescent bulbs to LED's, but could improved safety be another, less obvious benefit? The objective of this paper is to evaluate crashes at signalized urban intersections to determine whether or not crashes were reduced after the installation of LED traffic signals. A before-and-after analysis was conducted for eight intersections using empirical Bayes estimation. The analysis revealed an increase of about 71% in crashes after the installation of LED traffic signals.

254 Contributed Oral Poster Presentations: Section on Quality and Productivity

Section on Quality and Productivity

Monday, August 1, 2:00 p.m.–3:50 p.m.

Process Control When Items Are Subject To Misclassification

✦William S Griffith, University of Kentucky, Department of Statistics, University of Kentucky, Lexington, KY 40506, *william.griffith@uky.edu*; Michelle L. DePoy Smith, Eastern Kentucky University

Key Words: Process Control, Attribute, Acceptance Sampling, Misclassification, Runs

Various papers have considered the possibility of misclassification of items when they are inspected. In this paper, we consider online process control in which every hth item is inspected. The item is subjected to repeated independent classifications and will be ultimately judged to be a conforming item if there are k consecutive judgments that it is conforming prior to total of f judgments that it is nonconforming. When the item is ultimately judged to be nonconforming, the process is stopped and a search for a cause is conducted. If no cause is found the process is put back online. When the item is judged as conforming, the process continues. There is a possibility of misclassification of a conforming item as nonconforming and a nonconforming item as conforming. We will derive the probabilities of several quantities of interest for this model.

Performance Analysis Of A Mewma Controller For Mimo Processes Subject To Metrology Delay

✦Chien-Hua Lin, Providence University, 200 Chung-Chi Rd., Salu Dist.,, Taichung City 43301, Taiwan, Taichung, 43301 Taiwan, R.O.C., *chlin@pu.edu.tw*

Key Words: Run to Run (R2R) Process Control, MEWMA Controller, Stability Conditions, Metrology delay, Virtual Metrology

To maximize the competitiveness of semi-conductor manufacturers, they try to increase wafer sizes and reduce measurement devices. Metrology delay is a natural problem in the implementation of run-to-run process control scheme in semi-conductor manufacturing processes. Wu et al. (2008) demonstrated the influences of metrology delay on both the transient and asymptotic properties of the product quality when the single-input and single-output process is compensated by an exponentially weighted moving average (EWMA) controller. However, many semi-conductor manufacturing processes have multiple-input and multiple-output variables. To overcome this difficulty, Good and Qin (2006) proposed a multivariate EWMA controller to compensate it, and only discussed the stability conditions of a MEWMA controller. In practical applications, an optimal control scheme has more criteria than a stability region of discount matrix. In this talk, based on the criterion of minimizing asymptotic mean squares errors, we show that how to choose the optimal discount factor for various combinations of metrology delay. In addition, we discuss the ability of virtual metrology applied in delay MIMO processes.

Using Latent Variables To Estimate Parameters Of Inverse Gaussian Distribution Based On Time-Censored Wiener Degradation Data

✦Ming-Yung Lee, PROVIDENCE UNIVERSITY, 200 Chung Chi Rd., Taichung, 43301 Taiwan, *mylee@pu.edu.tw*

Key Words: inverse Gaussian distribution, latent variable, modified maximum likelihood estimator, Wiener process, time-censored

How to effectively assess the lifetime distribution of a highly reliability product during the product development stage in a timely manner is an important issue for a manufacturer. The traditional accelerated life testing, where samples are tested under higher stress to accelerate failures, may not be the most effective way in this context. On the other hand, if we could measure the degradationof a critical product characteristic over time, these measurements may provide useful information for estimating the lifetime without observing the failure of the product. Tseng, Tang, and Ku (2003) proposed a (time-transformed) Wiener process for the degradation of the brightness of LED bulbs in a scanner (CIS). Under their model, Lee and Tang (2007) proposed a modified EM algorithm to obtain modified MLEs (MMLEs) for both location and scale parameters of the lifetime distribution of LEDs. Due to time-censoring, the MMLE of the scale parameter was asymptotically biased. In this paper, we propose a method based on the latent variable technique to obtain a new estimate (LVE) of the scale parameter, and prove that this LVE is a consistent estimate with a smaller deviation than the MMLE.

The Effects Of Choice Of Housekeeping Gene On Efficiency In Real Time Pcr Data Analysis

✦Yi Guo, University of Florida, 1329 SW 16th Street, Room 5059, PO Box 100177, Gainesville, FL 32608, *yiguo@ufl.edu*; Michael Pennell, The Ohio State University; Soledad Fernandez, The Ohio State University; Dennis Pearl, The Ohio State University; Thomas Knobloch, The Ohio State University; Christopher Weghorst, The Ohio State University

Key Words: qPCR, Housekeeping genes, relative efficiency, paired t-tests

Real time polymerase chain reaction (quantitative PCR or Q-PCR), a highly sensitive method of quantitatively measuring gene expression, is widely used in biomedical research. To produce reliable results, it is essential to use stably expressed housekeeping genes (HKGs) for data normalization. In this study, a Q-PCR dataset that contains 12 target genes and 3 different HKGs was analyzed to examine the effect of different HKGs on the efficiency of paired t-tests. Our results showed that in addition to using stably expressed HKGs, it is also important to use HKGs with low levels of variability under the experimental conditions to increase the efficiency, and hence power of the t-test.

Evaluation Of A Process Change For Aluminum Laminate Production

✦Christin S Whitton, Los Alamos National Laboratory, NM 87544, *christin.s.whitton@gmail.com*; Leslie M. Moore, Los Alamos National Laboratory

Key Words: process change, Bayesian analysis

An experiment was conducted to evaluate a process change to production of laminated aluminum foil sheets. This poster will describe the proposed process change, the experiment plan, and the initial analyses of experiment results. Currently, a combination of TeflonÆ/fiberglass and polyester cloth is used to laminate the foil and the foil is laminated in a single sheet. The engineers in charge of production are considering process changes replacing the current materials with PacoThaneÆ and PacoPadÆ, laminating materials produced by PacoThane Technologies, and producing the foil in a double sheet. These process changes may reduce introduction of foreign material from the polyester cloth into the product and allow for double production. However, any process change must not affect the performance of the foil, measured by peel strength. The engineers conducted a small experiment to assess the possibility of changing the production process. We will present results and analyses of the experiment and discuss appropriate statistical assumptions, particularly statement of the null hypothesis, use of confidence intervals for evaluating process change, and Bayesian analysis methods.

Prediction Region As A Quality Measure For Dose-Response Curves

Zhibao Mi, VA Cooperative Studies Program; ✦Nan Song, Precision Therapeutics, Inc., 2516 Jane Street, Pittsburgh, PA 15203, *nansong04@gmail.com*; Kui Shen, Precision Therapeutics, Inc; Joseph Collins , VA Cooperative Studies Program; Gong Tang, Department of Biostatistics, University of Pittsburgh

Key Words: Dose-response curves, Prediction region, Quality control

Even though dose-response curves have been widely used as efficacy readouts in the life sciences, there is still interest in improving quality control methods for dose-response curves that do not belong to a parametric family. We propose constructing a prediction region based on a group of standard or previously accumulated dose-response curves as a quality control estimate of future generated dose-response curves with a predetermined level of probability. Lack of curve fitted parameters, such as LD50, the sample means, and the variances and covariances of the responses at various doses are used to build prediction regions using a multivariate technique. The prediction regions were built and tested using dose response data from human tumor cell growth curves inhibited by a panel of chemotherapy agents. If a testing dose response curve falls out the 95% prediction region, the curve is considered to be a failure, and the experimental condition is recalibrated. When compared to simultaneous prediction bands using the Student maximum modulus technique, the prediction region using the multivariate technique holds more stringent criteria for testing dose-response curves.

Relationships Between The T-Square Statistic And The Influence Function

✦Robert L. Mason, Southwest Research Institute, San Antonio, TX 78228-0510, *rmason@swri.org*; Youn-Min Chou, University of Texas at San Antonio; John C. Young, Retired

Key Words: correlation coefficient, influential observation, MYT decomposition

Hotelling's T-Square statistic has many applications in multivariate analysis. In particular, it can be used to measure the influence that a particular observation vector has on parameter estimation. For example, in the bivariate case, there exists a direct relationship between the ellipse generated using a T-Square statistic for a single observation and the hyperbola generated using Hampel's influence function for the corresponding correlation coefficient. In this paper, we jointly use the components of the T-Square statistic in the MYT decomposition and some influence functions to identify outliers or influential observations. Since the conditional components in the T-Square statistics are related to the possible changes in the correlation between a variable and a group of other variables, we consider the true influence functions of the correlations and multiple correlation coefficients. Two finite-sample versions of the true influence functions are used to find the estimated influence function values.

Statistical Process Control Via Times Series

✦Maria Emilia Camargo, Universidade de Santa Cruz do Sul, Av. Jo„o Machado Soares, 3199, Santa Maria, RS, Br, Santa Maria, 97110-000 Brazil, *kamargo@terra.com.br*; Angela Isabel dos Santos Dullius, Universidade Federal de Santa Maria; Walter Priesnitz Filho, CTISM-Universidade Federal de Santa Maria; Ivonne Maria Gassen, Universidade de Santa Cruz do Sul

Key Words: Quality Control, ARIMA Models, Correlated Data, Relative Efficiency

Technological development has reduced the variability items produced on large scale. Today a small change in the process can be critical, requiring rapid action to eliminate it. Traditionally, control charts are used to distinguish between the common causes of variation and the assignable causes of variation, to process generating independent and identically distributed random variables. In this paper the ARIMA

models to monitoring serially correlated data have been proposed. The results are then compared with those from traditional techniques: Classical Shewhart control charts and Cumulative Sum Method (CUSUM). The expected number and the expected time for some real and simulated series has been analyzed as well as the verification of relative efficiency between the methods.

Quality Assessment Study For Perception Failures Of The Customers

✦Berna Yazici, Anadolu University, Yunus Emre Kamp¸s¸, Fen Fak¸ltesi, Istatistik Bˆl¸m¸, Eskisehir, 26470 Turkey, *bbaloglu@anadolu.edu.tr*; Bet¸l Kan, Anadolu University; Ahmet Sezer, Anadolu University

Key Words: Quality Control, Failure of Perception, Production, Frequency

Some of the customers may perceive the problem they live about product as a fault. There is a big amount of warranty expenses for the big companies sourcing from the customers' failure perception. That expense increases regarding the service since all claims can not be solved by reading the user manuals or by the support of the call centers. I n this study, a real life data set that belongs to 2010 from a big white goods manufacturer company that exports products to more than 90 countries is examined. In this manner the data set in question is analyzed in order to reduce the warranty expenses sourcing from the perception failures of the consumers and the results are interpreted.

A Statistical Algorithm For Assessing Homogeneity Among Dried-Blood-Spot (Dbs) Pools

✦Maya Sternberg, CDC, 1600 Clitfon Rd, MS F 43, Atlanta, GA 30033, *mrs7@cdc.gov*; Victor De Jesus, CDC; Joanne Mei, CDC

Key Words: quality assurance, homogeneity, ANOVA, equivalence

Effective screening of newborns using dried-blood-spot (DBS) specimens collected at birth helps prevent mental retardation and premature death through early detection. Each year, millions of babies in the U.S. are routinely screened for certain genetic, endocrine, and metabolic disorders. CDC prepares and distributes more than 500,000 DBS per year to US domestic and international laboratories as part of a quality assurance program to ensure high levels of technical proficiency. Prior to distribution, the homogeneity of the DBS materials is assessed by randomly sampling DBS cards from across the entire pool, and punching disks from spots across each card. The natural statistical test to assess homogeneity is a one way random effects analysis of variance. However, instead of basing the homogeneity assessment on the standard F-test, Fearn and Thompson (2001) suggest a test called "sufficient homogeneity," which bases the decision on defining a permissible target variance and comparing that value to a one sided 95% confidence interval on the between-sample variance, (L, infinity). This abstract proposes a new homogeneity test based on principles of equivalence hypothesis tests.

A Robust Classifier For The Quality Control Of Reverse Phase Protein Array

✦Zhenlin Ju, UT MD Anderson Cancer Center, 1515 Holcombe Blvd, Houston, TX 77030, *zju@mdanderson.org*; Kevin Coombes, UT MD Anderson Cancer Center; Wenbin Liu, UT MD Anderson Cancer Center; Gordon Mills, UT MD Anderson Cancer Center; Paul Reobuck, UT MD Anderson Cancer Center; Siwak Doris, UT MD Anderson Cancer Center

Key Words: RPPA, Quality control, Bioinformatics, Proteomics

High-throughput reverse phase protein array (RPPA) technology enables measurements of the protein expression levels of thousands of samples in parallel. However, sample lysate preparation, slide printing, hybridization and washing may create substantial variability in the quality of the RPPA protein expression data. Thus far, there has not been an algorithm available for RPPA data quality control (QC). Therefore, we have developed a novel classifier for the quality control of RPPA experiments using a logistic regression model (LRM). The outcome of the prediction model is the probability that a slide is of good slide over the probability that a slide is of poor slide, ranging from 0 to 1. We conclude that the proposed classifier outlined in this study can sufficiently distinguish good quality from poor quality data, and can be used to retain good quality slides so that the normalization schemes protein expression patterns, and advanced biological analyses will not be drastically impacted by erroneous measurements and systematic variations.

Determination Of The Time And Expected Number Of Inspections In A Manufacture Of Agricultural Tractors

Maria Emilia Camargo, Universidade de Santa Cruz do Sul; ✦Walter Priesnitz Filho, CTISM-Universidade Federal de Santa Maria, , *walterpf@gmail.com*; Angela Isabel dos Santos Dullius, Universidade Federal de Santa Maria; Pelayo Olea Munhoz, Universidade de Caxias do Sul; Stephanie Russo Fabris, Federal University of Sergipe

Key Words: Shewhart Charts, Monitoring, Cusum, Bayesian Model

Statistical process control (SPC) techniques are widely used in industry for process monitoring and quality improvement. Statistical process control (SPC) techniques are widely used in industry for process monitoring and quality improvement. Tradicional SPC charts are based on a fundamental assumption that process data are statistically independent and normally distributed when the process is in control. In this paper, data have been collected and analyzed from an industry specializing in the manufacture of agricultural tractors using the classical Shewhart control charts, cumulative sum method, and bayesian model searching to determine the time and expected number of inspections between the occurrence of out of control situations and their detection.The result of this research will come to contribute still more, with information of great importance for the taking of decision of local industries, therefore it is a form of doing continuous monitoration, making possible an improvement in the control of the system.

255 Contributed Oral Poster Presentations: Section on Statistical Graphics

Section on Statistical Graphics
Monday, August 1, 2:00 p.m.–3:50 p.m.

Objective Identification Of Extreme-Most Daily Maximum/Minimum And Hour-To-Hour Historical Temperature Patterns Using Principal Components Analysis

✦Charles Fisk, U.S. Government, 91320, *cjfisk@att.net*

Key Words: Temperature Patterns, Principal Components Analysis, First Component Correlation Loadings, Floating Bar graphs, First Component Covariance Loadings

A familiar method of depicting day-to-day max/min temperature data is the floating-bar or "hi-lo" chart. In viewing individual charts, the question may arise on how typical the patterns are relative to those of other years covering the same calendar segment, and to this end, it would be useful to have objective, quantitative means of characterization and comparison. The same could apply to line-depicted day-to-day hourly temperature observations. Utilizing Downtown Los Angeles daily max/min (1921-2010) and Los Angeles International Airport hourly (1939-2010) temperature data, the utility of Principal Components Analysis (Correlation and Covariance, each unrotated) is demonstrated for this purpose. First component correlation loadings characterize "shape", first covariance loadings, "spread". First component correlation scores are perfectly correlated with climatological means for the day/hour and temperature type in question. The highest and lowest correlation/covariance loadings identify the most anomalous patterns in terms of these attributes, and for a hierarchy of calendar periods, graphs of the most extreme configurations are presented as illustrative examples.

The Impacts Of Bp Oil Spill On Fish And Birds

✦Haolai Jiang, Western Michigan University, 1014 Claymoor Dr, Apt 2D, Kalamazoo, MI 49009, *xiaoflyingbear@gmail.com*

Key Words: oil spill, fish, birds, evironment, graphics

BP oil spill is a huge shock to the environment of United State. Thousands of wild animals dead by the effect. This poster will show the impacts graphically. The specific focus is on the living condition changes of fish and birds.

Detecting Oil Through Indirect Measurement

✦Patrick Robert Steele, The College of William and Mary, CSU 2936, PO Box 8793, Williamsburg, VA 23187, *prsteele@email.wm.edu*; Jennifer Anne Thorne, The College of William and Mary; Amy Olivia Russell, The College of William and Mary

Key Words: optimization, sampling

There are many difficulties associated with measuring oil contamination in the ocean. Not only are specialized equipment and procedures necessary to collect and analyze the data, the collection itself can expose those involved to hazardous environments. We explore data collection methods designed to minimize the above difficulties; namely, we consider: minimizing the number of observations required to accurately gauge the extent of the oil spill; developing collection techniques that allow for indirect measurement of the oil spill; and maximizing the data collected given fixed financial resources.

Visualizing Of Clusters Of Earth Grid Cells Based On Global Atmosphere Data And Cluster Changes Over Time

✦Daniel Carr, George Mason University, Statistics MS4A7, George Mason University, Fairfax, VA 22030, *dcarr@gmu.edu*; John Ashley, NVidia

Key Words: Evolving clusters, color palettes, land and ocean

Studying global multivariate multi-altitude atmosphere data derived from satellite monitoring poses visualization challenges. Based on radiances obtained from the Atmospheric Infrared Spectrometer on the Aqua satellite, NASA computes geophysical parameters for satellite foot prints. One level 3 NASA product consists of monthly entropy constrained vector quantized summaries of footprints in 5 degrees earth grid cells. The number of summary vectors varies from cell to cell. The Wasserstein metric provide a way to obtain the expected distance between grid cells over space and time. Clustering based on very large distance matrices provide an overview based on clusters that grow or shrink and change locations over time. Encoding many clusters using color in a readily understandable way is a visualization challenge. The proposed solution uses sea surface temperature of clusters primarily over the ocean as a basis for picking colors from an ocean color palette. Similarly, for land clusters the NDVI greenness index proves a basis for picking colors from a land palette. for primarily land clusters. The graphics also address change blindness is showing changes across sessions and years.

Dynamic Graphics For Exploring And Understanding Multiple Regression Diagnostics

✦Mervyn G Marasinghe, Iowa State University, Department of Statistics, Snedecor Hall, Ames, IA 50011, *mervyn@iastate.edu*; Christopher Bruno, Iowa State University

Key Words: dynamic graphics, regression diagnostics, regression graphics, multicollinearity, model selection, interactive tools

Regression techniques are extremely popular across many disciplines. Modern statistical software systems incorporate many of the diagnostics and static graphical tools currently available. However, practitioners still have to rely on conventional methods of examining several models and comparing the relevant diagnostics manually for making decisions such as model selection, analysis of multicollinearity or determining the effect of cases on model fit. In this paper, we describe several tools that incorporate well-known dynamic graphic techniques, that allows the user to examine and compare model fits using several criteria and eliminates the element of artificiality in model selection.

256 Contributed Oral Poster Presentations: Section on Statistical Learning and Data Mining

Section on Statistical Learning and Data Mining
Monday, August 1, 2:00 p.m.–3:50 p.m.

Uncovering Patterns Of Technology Use In Consumer Health Informatics

✦Man Hung, University of Utah, 590 Wakara Way, Salt Lake City, UT 84108, *man.hung@hsc.utah.edu*

Key Words: electronic medical record, consumer health informatics, data mining, patient health record, technology adaption, bioinfomatics

The concept of electronic medical record represents an evolving aspect of consumer health informatics. The patient portal is most often considered to be a patient view to an electronic medical record that is owned by the clinician or health care system, where the personal health record is an individual-controlled storehouse of health information. The present study is an evaluation of part of a larger project that integrates personal health records with electronic medical records, which called the Unified Health Resource (UHR). This study aims to examine patients' adaptation of the UHR. Data mining is used to explore the association between technology use and various patient and system characteristics.

Data Preprocessing And Variable Selection In The Study Of Proteomic Mass Spectrometry

✦Chamont Wang, The College of New Jersey, , *wang@tcnj.edu*; Charlene Wang, HealthFirst Inc.; Michele Meisner, The College of New Jersey

Key Words: Variable Selection, Data Preprocessing, False Discover Rate, Decision Tree, Stochastic Gradient, Regression

This study investigates a set of proteomic data, collected from the records of 216 individuals: 121 of those with cancer and 95 healthy volunteers. For each individual, there are 368,749 pieces of the spectra in the raw data. In our investigation, we use a technique of Dynamic Binning to merge adjacent spectra by assigning similar compounds to the same spectrum without minimizing peak resolution. The process reduced the raw data from 1.16 Gb to 9.3Mb in 5,155 bins. Our study compares the effect of this technique with other types of binning. Within each bin, one can take mean, max, SD, moving average and other types of statistics for predictive modeling. This study compares the efficiencies of these statistics in the prediction of cancer patients. Furthermore, the study investigates the effects of Variable Selection via False Discover Rate as discussed in Efron (2010, 2008) and Benjamini and Hockberg (1995). In addition, we used various techniques from Dudoit, Shaffer, and Boldrick (2003). We compare these results with the variables selected by Decision Tree, Stochastic Gradient, Regression, and Partial Least Squares.

Multiple Kernel Learning Classification Of Lupus Disease From Structural Mri

✦Cen Guo, University of Michigan, 439 West Hall, 1085 South University, Ann Arbor, MI 48109, *gcen@umich.edu*

Key Words: MRI, Multiple Kernel Learning

Structural Magnetic Resonance imaging (MRI) was used as in the clinical trails extensively in recent years. Classification analysis of anatomical MRI can help in diagnosis of disease. Traditional classification analysis using svm method is either based on whole-brain image with tens of thousands of variables or a particular region of interest which needs extra information about the disease to specify. In this work, we proposed a hierarchical method to reduce the dimension and select significant regions automatically. The first step is to train a svm model for every small region across the whole brain. In the second step, a multiple kernel learning scheme is applied to significant regions to further improve the performance and select important features simultaneously. A real data analysis of Lupus disease shows that this new method can outperform the traditional one step svm method.

Principal Component And Independent Component Analysis Of Brain-Generated Biopotencial Measurements

✦George Freitas von Borries, Universidade de Brasilia, Brasilia - DF, International 70910900 Brazil, *gborries@unb.br*; Murilo Coutinho Silva, Universidade de BrasÌlia; Loyane Christina Soares Rocha, Universidade de BrasÌlia; Ricardo Freitas von Borries, University of Texas at El Paso

Key Words: Principal Component Analysis, Independent Component Analysis, Electroencephalography

Recent techniques in digital signal processing allow detection of different kinds of mental activity from brain-generated biopotentials. Electroencephalograph (EEG) methods, as opposed to invasive measurements, have the advantage of requiring only surface biopotentials to be acquired, and are therefore more suitable for regular-use HMI equipment in applications as classification of thought patterns for machine control by persons with motor disabilities. This paper uses Principal Component Analysis (PCA) and Independent Component Analysis (ICA) of EEG signals to investigate data reduction source activated areas of brain when submitted to a sequence of visual stimuli. The data was collected in the Biopotentials Imaging Lab (BIML) at UTEP, using a 128-electrode acquisition system. EEG data reduction and source identification allow one to understand brain activity during different tasks and to reduce the number of dimensions in classification procedures. By applying this technique, we expect to improve classification of different tasks with respect to usual classification techniques applied to raw data and to better understand brain activation by visual stimuli.

Multi-Label Classification Via Binary Markov Networks

✦Jie Cheng, University of Michigan, 1085 S University Ave, 439 West Hall, Ann Arbor, MI 48109, *jieche@umich.edu*

Key Words: multi-label classification, binary markov networks, pseudo likelihood

Multi-label classification refers to the scenario in classification that each instance is associated with a subset of labels rather than one. The labels are not mutually exclusive and often correlated. In this project, we first propose to transform multi-label classification into a multivariate binary regression problem. Then we introduce an Ising model with covariates to explicitly model the conditional distribution of the class labels given the covariates. Pseudo-likelihood is adopted to develop a computationally efficient estimation procedure. We also investigate the choice of evaluation measures in connection to different prediction rules, which is further illustrated by numerical studies. The proposed method is applied to a popular Yeast dataset and shows promising result.

Exploring Genetic Risk For Breast Cancer Using An Ensemble Of Tree-Based Classifier

✦Bethany Wolf, Medical University of South Carolina, 135 Cannon Street, Suite 303, Charleston, SC 29425, *wolfb@musc.edu*; Elizabeth Hill, Medical University of South Carolina; Elizabeth H Slate, Medical University of South Carolina; Carola Neumann, Medical University of South Carolina; Emily Kistner-Griffin, Medical University of South Carolina

Key Words: esemble classifiers, binary predictors, predictor interactions

The predictive accuracy of tree-based classifiers (learners) can be improved by using an ensemble of learners to predict an observation's class. Ensembles allow averaging across weak learners (unbiased classifiers that are highly variable) resulting in an unbiased aggregated learner with reduced variability. Some ensemble methods also provide measures of predictor importance, allowing scientists to discover potential biological markers predictive of disease. For example, single nucleotide polymorphisms (SNPs) are thought to alter risk of developing disease or prognosis once disease occurs. Measures of predictor importance from an ensemble enable ranking of SNPs and SNP interactions according to association with disease outcome. We examine the performance of several ensemble methods in simulation studies for predictive capability and for ability to correctly identify binary predictors and predictor interactions associated with a binary outcome. We apply these methods to a subset of data from the Cancer Genetic Markers of Susceptibility (CGEMS) Breast Cancer Scan which includes SNPs from select genes and explore associations between the SNPs on these genes and breast cancer.

Finding Biomarkers Of Lung Cancer Using Different Classification Methods

✦Jianghong Deng, George Mason University, 10900 University Boulevard MS 4E3, Manassas, VA 20110 USA, *jdeng@gmu.edu*; Angela Zupa, I.R.C.C.S.CROB Centro di Riferimento Oncologico della Basilicata; Giuseppina Improta, I.R.C.C.S.CROB Centro di Riferimento Oncologico della Basilicata; Alessandra Silvestri, Division of Surgical Oncology, CRO-IRCCS,National Cancer Institute; Michele Aieta, I.R.C.C.S.CROB Centro di Riferimento Oncologico della Basilicata; Pellegrino Musto, I.R.C.C.S.CROB Centro di Riferimento Oncologico della Basilicata; Lance Liotta, George Mason University; Julie Wulfkuhle, George Mason University; Emanuel F. Petricoin, George Mason University

Key Words: Discriminant Analysis, Support Vector Machine, K-nearest Neighbor, Random Forest, Biomarker

Discriminant Analysis, Support Vector Machine, K-nearest Neighbor, and Random Forest are proposed to identify protein biomarkers of lung cancer dataset. K-Fold Cross-Validation for these classification methods will be performed. Accuracy rates of classification will be used to measure the performance among these four methods.

Using A Tree-Structured Survival Method To Detect Drug-Drug Interactions

✦Liping Huang, Hoffmann La-Roche, 08837 USA, *lipinghuang89@gmail.com*

Key Words: Survival Tree, Drug-drug Interaction

Background: A tree-structured survival (TS) method been increasingly useful in drug safety. However, the TS method has been rarely subjected to an examination of their adequacy. Objectives: The aim is to compare the predictive power of TS in detecting interactions across different sample sizes and proportions of censoring. Methods: Based on the combinations of sample sizes (500 and 3000) and three different proportions of censoring (50%, 70% and 90%), 500 simulated data sets were created for each combination. Each simulated data set contained thee types of predictors and their interactions with known true coefficients. Corresponding numbers of survival trees were created based on each simulated data set of a combination to examine power of detecting direct and indirect interactions. In addition, a real world example was applied to examine drug-drug interactions among a diabetic patient population who encountered a cardiovascular event. Results: By using the TS method, the power of detecting interactions for data with lower or median percentage of censoring is higher than the data with high percentages of censoring. This is also true for data with large sample size.

Dimension Reduction For Classification Using M-Method Techniques

✦John Beeson, Baylor University, , *john_beeson@baylor.edu*

There have been numerous papers dedicated to the dimension reduction of microarray data using Partial Least Squares and Principle Components techniques. We propose using a dimension reduction method that uses the sample means and covariance matrices of the microarray data, so that the dimensionality can be reduced with minimal loss of classification information.

257 Contributed Oral Poster Presentations: Section on Statistics and Marketing

Section on Statistics and Marketing

Monday, August 1, 2:00 p.m.–3:50 p.m.

Socio-Economic CUM Cultural Characteristics and Choice of Telecoms Business Network

✦Hamadu Dallah, University of Lagos, Department of Actuarial Science and Insurance, Faculty of Business Administration, Lagos, International 1014 Nigeria, *dallaram2007@yahoo.com*; Charles Omoera, University of Lagos; Rasaki Bakare, University of Lagos; Ismaila Adeleke, University of Lagos; Benjamin Oghojafor, University of Lagos

Key Words: Telecoms Business Network, Consumers' preference, market survey, Poisson regression, government policies, Nigerian business environment

The development of telecoms sector has experienced a major process of transformation in terms of its growth, technological content and market structure in recent times. The paper investigates the effect of socio-economic and cultural characteristics on telecoms consumer's network preference using the generalized Poisson regression model. A market survey was conducted to gauge the enabling data needed in the present study. Findings of this research will assist the industry in designing suitable products to meet the expectations of consumers in Nigerian business environment. Finally, the research also discusses government enabling policies put in place for good service delivery in telecoms industry.

Socio-Economic CUM Cultural Characteristics and Choice of Telecoms Business Network

✦Hamadu Dallah, University of Lagos, Department of Actuarial Science and Insurance, Faculty of Business Administration, Lagos, International 1014 Nigeria, *dallaram2007@yahoo.com*; Ismaila Adeleke, University of Lagos; Charles Omoera, University of Lagos; Rasaki Bakare, University of Lagos; Benjamin Oghojafor, University of Lagos

Key Words: Telecoms Business Network, Consumers' preference, market survey, market survey, government policies, , Nigerian business environment

The development of telecoms sector has experienced a major process of transformation in terms of its growth, technological content and market structure in recent times. The paper investigates the effect of socio-economic and cultural characteristics on telecoms consumer's network preference using the generalized Poisson regression model. A market survey was conducted to gauge the enabling data needed in the present study. Findings of this research will assist the industry in designing suitable products to meet the expectations of consumers in Nigerian business environment. Finally, the research also discusses government enabling policies put in place for good service delivery in telecoms industry.

Customer Segmentation And Retention In Nigerian Telecommunication Service

✦Rasaki Bakare, University of Lagos, Department of Business Administration, Faculty of Business Administration, Lagos, International 1014 Nigeria, *adebak2@yahoo.com*; Charles Omoera, University of Lagos; Ismaila Adeleke, University of Lagos; Hamadu Dallah, University of Lagos; Benjamin Oghojafor, University of Lagos

Key Words: teleCommunication Industry, Marketing Tactics, Consumer Characteristics, Cluster Analysis, Customer Retention

The landscape of the telecommunication industry in Nigeria changed drastically since the deregulation of telecommunication sector in early 2000. Number of service providers has increased from one state monopoly service provider, to about 10 within in the last decade. The increase in number of service providers and subscribers coupled with stiff competition in the industry informed the need for providers to strategize on marketing tactics. This article examined the relationship of service delivery, telecom market segmentation and consumers' profiling on customers' retention. Cluster analysis was used to achieve market segmentation of consumers' characteristics on various networks. The level of contributions of the explanatory variables on retention on a network is analyzed. The findings from the study can be used by service providers to develop innovative products and services to meet the ever-changing needs of existing customers and to attract new customers in Nigerian business environment. Some policy implications are also discussed.

Sales Forecast Using Classical Structural Model And Artificial Neural Network

✦Maria Emilia Camargo, Universidade de Santa Cruz do Sul, Av. Jo„o Machado Soares, 3199, Santa Maria, RS, Br, Santa Maria, 97110-000 Brazil, *kamargo@terra.com.br*; Walter Priesnitz Filho, CTISM-Universidade Federal de Santa Maria; Angela Isabel dos Santos Dullius, Universidade Federal de Santa Maria; Suzana Leit„o Russo, Federal University of Sergipe; Marta Elisete Ventura da Motta, Universidade de Caxias do Sul

Key Words: Classical Structural Model, Sales, Neural Network, Forecast, Intervention

In this article we analyzed the sales of a large size enterprise located in the state of Santa Catarina, Brazil, for the period January 2000 to December 2010, using the Classical Structural Model with intervention and artificial neural networks. The structural model with interventions presented a residual variation of 0.0009, where as the neural network model presented a residual variation of 0.0004. The chosen neural network presented, for the last 12 months, better forecasts than that structural model . The mean absolute error of the forecast was of 3.4782 to neural network model and of 5.4789 to structural model with interventions. The model obtained by the Neural Network was superior to Classical Structural Model, in adjustment as well as in forecasting for the data analyzed.

258 Contributed Oral Poster Presentations: Section on Statistics in Sports

Section on Statistics in Sports

Monday, August 1, 2:00 p.m.–3:50 p.m.

A 10-Year Study of the Relative Importance of Driving Accuracy and Driving Distance on Scoring Average on the PGA Tour

✦Erik Heiny, Utah Valley University, Mathematics, Orem, UT 84058 United States, *erik.heiny@uvu.edu*; Robert Heiny, University of Northern Colorado

Key Words: PGA Tour, regression

This study examines the relative importance of driving skills (driving distance-DD and driving accuracy-DA) on the PGA Tour over the last 10 years. The dependent variable is scoring average. Static regression models are run for the years 2000-2009. Results of these regressions allow reasonable percentiles to be calculated for the two driving variables to compare the effect each has on scoring averages. Confidence intervals are constructed for both scoring average and scoring rank to assess the importance of the driving variables. The results suggest that driving accuracy is more likely to yield better results when trying to improve scoring rank.

Simulating Rare Baseball Events Using Monte Carlo Methods In Excel And R

✦Scott Nestler, Naval Postgraduate School, 2993 Bird Rock Rd., Pebble Beach, CA 93953, *stnestle@nps.edu*; Scott Billie, U.S. Military Academy at West Point; Michael Huber, Muhlenberg College; Gabriel Costa, U.S. Military Academy at WEst Point

Key Words: Baseball, Statistics, Simulation, Monte Carlo

Statistics and baseball have had a long and venerable relationship throughout the history of the sport. Practically anything and everything has been recorded and tracked by avid baseball fans. The same fans debate statistics and argue which of the hallowed records will ever be broken. Websites such as Baseball-Reference.com and Baseball-statistics.com now offer data sets that include hitting, pitching, fielding, and base running events for both individual players and collective teams. Simulating certain events, such as hitting streaks or number of wins in a season, have become effective approaches to answering "Will this record ever be broken?" One such seemingly-unbreakable record is Joe's DiMaggio's famous 56-game hitting streak of 1941. In this paper we apply Monte Carlo simulation techniques in both Microsoft Excel and the open-source statistical package R to determine the likelihood of such a rare event using actual data from the DiMaggio streak.

Implementing Random Forests For Hockey Hall Of Fame Induction Prediction: Applications To Language-Based Discrimination

Brian M Mills, University of Michigan Department of Sport Management; ✦Steven Salaga, University of Michigan Department of Sport Management, 1402 Washington Heights, Ann Arbor, MI 48109, *salaga@umich.edu*

Key Words: Random Forest, Hockey, Classification, Multidimensional Scaling, Logisitic Regression, Sports

Discrimination based on race, ethnicity, age, religion or other factors can significantly affect the labor market and earnings for those individuals whom may be discriminated against. Previous work in the labor market of professional hockey has found evidence of discrimination based on language, specifically against French speaking professional hockey players in English hockey leagues. We extend this research to the subjective voting involved in Hockey Hall of Fame inductions. We first predict current National Hockey League player induction into the Hockey Hall of Fame using the very competitive classification technique, Random Forests, using training data from historical NHL statistical records and previous inductions. Secondly, we extend on previous labor investigations to consider the possibility of induction exclusions of French speaking professional hockey players. We conclude with multidimensional scaling visualizations and rescale the Random Forest votes to induction probabilities using a logistic regression. Preliminary results show little evidence for language-based discrimination and promising prediction error for classification in sports.

Training In The Off-Season: Do Elite Gyms Improve Performance In Baseball?

✦Isabel Elaine Allen, Babson College, Arthur M. Blank Center for Entrepreneurship, Wellesley, MA 02457, *allenie@babson.edu*; Julia Seaman, UCSF

Key Words: baseball, sports, training, time series, prediction

Off-season training in baseball, and in all professional sports, has changed greatly over the past 25 years. Prior to the huge salaries, players needed to obtain a second job in the off-season to make ends meet. Early training involved little aerobic or weight training and was likely to only encompass isometrics. Training has seen a progression from aerobics and weight training to private trainers and now to all-inclusive elite gyms. Do these gyms, which include special diet and nutrition programs as well as physical training to strengthen the 'core' and specific programs tailored to each athlete, live up to their promise of better performance and fewer injuries? With data from the last 8 years of training from an elite gym, we can examine overall trends in performance and fitness and examine major league baseball players before and after their off-season training and begin to test whether the gyms live up to their claims and predict future performance gains (or not) of attendees.

The Effects Of Race On Called Strikes And Balls

Jeff Hamrick, Rhodes College; ✦John Rasp, Stetson University, DIS Department, Unit 8398, 421 N Woodland Blvd, DeLand, FL 32723, *jrasp@stetson.edu*

Key Words: baseball, sports, racial discrimination, race, logit, logistic

In the years since Jackie Robinson broke the "color barrier" in baseball, overt racism has become socially unacceptable in mainstream American society. However, more subtle patterns of discrimination persist. Baseball provides a convenient laboratory for examining the extent to which discriminatory behaviors remain, even at the subconscious level. We examine pitch-by-pitch data for the 1989 through 2010 major league baseball seasons. We focus on the umpire's subjective decision as to whether a pitch is called a ball or strike. We find that the pitcher's and hitter's race has a small, but statistically significant, effect upon the outcome of that call.

Analyzing Game Flow In The Nfl

✦Luis Fernando Campos, University of California, Berkeley, 2914 A Deakin St, Berkeley, CA 95705, *lfcampos@berkeley.edu*; Miklos Zoltan Racz, University of California, Berkeley

Key Words: Sports, football, NFL, game flow

We have collected detailed play-by-play data for every game in the 2010 season of the professional American football league (NFL). The information focuses on offensive play-calling aspects of the game (passes, rushes, etc.) and also contains details of how each drive ends (touchdown, interception, etc.). Our goal is to analyze the decision-making process of teams in the NFL by treating each match as a sequence of plays instead of simply analyzing static observations (e.g. pass/run ratio). Can we predict the next play given the current state of the drive and past plays? Are team tactics significantly different across the league? How much does play-calling account for team success? The detailed nature of the data will provide us with the capability to answer these questions. We hope that this type of game flow analysis will provide us with additional insight into what makes a team successful.

A Logistic Model On Making The Nhl Postseason; Effort Defines Elite Teams

✦Mark Atwood, University of Massachusetts, Lowell, 1 University Ave., Lowell, MA 01854, *MarkAtwood1984@gmail.com*

Key Words: Logistic, NHL, Hockey, Playoffs, Sports

Although many factors contribute to an NHL team reaching postseason play, none are more so than points earned and goals allowed during the regular season. Using team-level data obtained from Hockey-Reference.com for the 2000-2001 season through 2009-2010 season and a logistic model to predict playoff appearances based on a number of factors, points were most significant ($p < 0.001$) and goals against, second ($p=0.0165$). All other factors entered into the model proved insignificant. Results from the model indicate that a team has the potential to increase the probability (OR [95% CI]) of making the postseason by as much as 70.9% (1.709 [1.411, 2.069]) by earning a single extra point. In contrast, although impractical, being able to allow 2 fewer goals increases the probability of making the playoffs by a maximum of 6.4% (0.968 [0.941, 0.995]). Players/coaches alike should note the importance of not being content with getting to overtime during the regular season, but rather strive to take that extra point. When playing a game that is clearly going to be a loss, it behooves the team to continue to try, rather than allow it to become a laugher.

Home Field Advantage And Toss In Limited-Overs Cricket

✦Ananda Bandulasiri Manage, Sam Houston State University, Dept. of Mathematics & Statistics, SHSU, Huntsville, TX 77340, *wxb001@shsu.edu*; Christina Kardatzke, Sam Houston State University

Key Words: cricket, home field advantage, logistics regression, quantile regression, winning the toss

There are several studies in the literature that have shown the significance of the home field advantage towards the outcome of limited-overs cricket matches. Some of these studies have also shown that the winning the toss does not give any significant advantage for the victory of matches. Authors will show some surprising new findings related to the significance of the home field advantage and winning the toss by analyzing limited-overs cricket data separately for "Day" and "Day/Night" matches. Quantile regression and logistic regression will be applied on the data obtained from previous matches for model fitting.

259 Contributed Oral Poster Presentations: Section on Teaching of Statistics in the Health Sciences

Section on Teaching of Statistics in the Health Sciences

Monday, August 1, 2:00 p.m.–3:50 p.m.

Meta Analysis For Health Care Data

✦Fanglong Dong, university of kansas school of medicine-wichita, 1010 N Kansas ST, Wichita, KS 67214, *fdong@kumc.edu*

Key Words: meta analysis, health care data, arrow plot

BACKGROUND: Meta-analyses of interventions in health care traditionally study dichotomous outcomes such as mortality. Increasingly, medical interventions are titrated to a surrogate outcome that is a continuous variable, such as a measure of cholesterol or glucose control. Meta-regression has correlated change in the clinical outcome based on change of the continuous surrogate by pooling outcomes in the intervention groups of trials, but this method ignores available data. METHODS: We have developed two dimensional meta-analytical plots of clinical outcome by surrogate outcome. For each trial an arrow leads from the results of the control group to the results of the intervention group. Dashed arrows indicate insignificant results. RESULTS: The change in the y-coordinate of each arrow is the relative outcome reduction. The slope of the arrow is the relation between the surrogate and clinical outcomes; a vertical or horizontal arrow indicates no relationship. This plot allows testing heterogeneity of slopes and modeling nonlinear relationships. CONCLUSION: We propose that two-dimensional arrow plots increase the information garnered from meta-analysis.

Cross-Cultural Comparisons Of Attitudes Towards Statistics In Health Science Students

✦Heibatollah Baghi, George Mason University, 4400 University DR., MS 5B7, Fairfax, VA 22030, *hbaghi@gmu.edu*; Melanie L Kornides, George Mason University

Key Words: Attitudes towards Statistics, Cross-Cultural Differences, Healthcare Statistics

The purpose of this study was to examine cross-cultural differences in attitudes towards statistics among graduate students in health science programs at a major U.S. university. Pre/post measures of attitude were administered to 165 students before and after 10 weeks of instruction in statistics to measure students' improvement. Of the participants, 82 were U.S. citizens and 83 were international students. Students were asked 25 questions (using a 5-point Likert scale) to measure self-perceived knowledge and attitudes regarding six statistical domains: (1) research design, (2) statistical computation, (3) statistical application, (4) statistical interpretation, (5) utility (the extent to which the person believes that statistical tools are useful in his/her work, and (6) self-confidence in the use of statistics. Our results show that health science students have, in general, a positive attitude toward statistics in all identified domains, although U.S. students value it more significantly in terms of utilities. A high correlation between the scores obtained from the attitudes scale and the test of statistical proficiency provides evidence of the criterion-related validity of the

A Teaching Experience In A Medical Department: From Concepts To Computation

✦Sunil Kumar Dhar, New Jersey Institute of Technology, 323 Martin Luther King Junior Boulevard, Newark, NJ 07102, *dhar@njit.edu*

Communicating statistics so that scientists can carry out their measurements collect data according to a designed experiment and become proficient in statistical computation was the main goal of a course taught in a medical department. This work shares the experience of teaching biostatistics to medical doctors, lab technicians, and medical students in a department of cell biology and molecular medicine, who had a minimal or rusty statistics background. The topics that were covered and the computational resources that were used, together with examples, will also be discussed.

Evaluation Of Supplementary Multimedia Aids To Improve Graduate Nursing Students' Attitude Toward Statistics

✦MyoungJin Kim, Illinois State University, IL 61704, *mkim2@ilstu.edu*

Key Words: Attitude toward statistics, Multimedia aids, Nursing students

While many nursing students find statistics difficult, the use of research and practice has been integrated in nursing for a long time. Currently, there is a lack of literature on statistical learning for graduate nursing students. The purpose of this descriptive mixed method study is to characterize the change of students' attitude toward statistics in a graduate nursing program from the beginning to course completion, evaluating the supplementary multimedia aids created by a researcher. A volunteer sample of 35 graduate nursing students will be recruited from a 4-year suburban university. Survey of Attitudes Toward Statistics (SATS) will be used to measure affect, cognitive competence, value, difficulty, interest and efforts. Data will be analyzed using descriptive statistics, correlations, and dependent sample t-tests. Understanding attitudes toward statistics in nursing students may assist in producing competent nurse researchers. Improved attitudes toward statistics will increase students' understanding of the relevance of statistics in nursing. Therefore, they will have the ability to be better doers and consumers of research for the best practice.

Effective Course, Lecture, Lecturing, And Lecturer

✦Martina Pavlicova, Columbia University, 722 West 168th street, rm. 637, New York, NY 10032, *pavlicov@gmail.com*; Phoebe Luong, Columbia University

Key Words: Teaching, Statistics, Biostatistics, preparation, Lecturing

The poster will summarize the components, which are necessary for the preparation of an effective course, in particular biostatistics for master public health students. The effective course is not only based on an effective structure but also has to incorporate an effectively designed lecture. We discuss in details the components of an effective lecture and make generalizable recommendations. We also discuss effective delivery and lecturing and suggest how to utilize concrete tips. Last but not least, we focus on the lecturer and give recommendations on what makes a lecturer an effective and memorable teacher. We believe that our poster is not only useful to PhD students preparing for their career in academia, but many seasoned lecturers can also benefit from our effective methods analyses.

Coping With A Wider Range Of Outcomes: Demonstration Of The Generalized Linear Model Platform In Spss Statistics

✦Letao Sun, Department of Educational Policy Studies and Evaluation, University of Kentucky, 131 Taylor Education Bd., Lexington, KY 40506, *letao.sun@uky.edu*; Hongwei Yang, Department of Educational Policy Studies and Evaluation, University of Kentucky

Key Words: regression, generalize linear model, link function, SPSS, multilevel

A common problem shared by researchers using regression analysis in health, social and behavioral sciences is that the outcome fails to be normally distributed. This includes data that are continuous, discrete, censored, and a combination of all these types. To appropriately analyze such data, link functions are utilized to link the mean of the outcome to the linear predictor, a linear combination of all predictor variables. Such a modeling framework is known as generalized linear modeling (GLM). Featuring a user-friendly environment, SPSS provides a procedure (GENLIN) for GLM analysis with a variety of built-in link functions. This procedure enables applied researchers to harness the power of GLM without any syntax-based coding. This study examines link functions available in the SPSS GLM procedure with a primary focus on how to quickly help applied researchers get started in using the procedure to tackle a wider range of data with different types of outcomes. Examples of applications in health sciences are presented. A future expansion of this study is to analyze more complex data structures (multilevel, etc.) in the GLM framework using this cost-effective, easy-to-use program.

260 ASA President's Invited Address

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

Monday, August 1, 4:00 p.m.–5:50 p.m.

Statistical Analysis: Current Position And Future Prospects

✦Sir David R. Cox, Nuffield College, Oxford, United Kingdom, *david.cox@nuffield.ox.ac.uk*

The importance and interest of statistics stems in large part from the wide range of applications of statistical ideas across the natural and social sciences and beyond. A major challenge is to preserve some unity of ideas while recognizing the genuine distinctiveness of individual applications. Illustrations will be given from a number of fields of application and the broad principles involved discussed, including the impact of computational developments. The role of foundational issues will be briefly mentioned.

261 IMS Presidential Address

IMS, International Chinese Statistical Association, SSC

Monday, August 1, 8:00 p.m.–9:30 p.m.

IMS Presidential Address

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

To general audience