
ENAR
Program and Abstracts
International Biometric Society
Eastern North American Region
March 20 – 23, 2005
Hilton Austin
Austin, TX
Spanning the Breadth of Biometrics:
From Ecosystems to Healthcare Systems
Spring Meeting with IMS and Sections of ASA

ENAR

Table of Contents

ENAR

Acknowledgements

EXHIBITORS

Blackwell Publishing, Inc.
Cambridge University Press
The Cambridge Group, Ltd.
CRC Press – Taylor and Francis Group
Cytel Software Corporation
Duxbury, Thomson
Insightful Corporation
Oxford University Press
Pfizer Global Research and Development
PPD, Inc.
Salford Systems
SAS Publishing
SIAM (Society for Industrial and Applied Mathematics)
Smith Hanley Associates LLC
Springer
StatSoft, Inc.
John Wiley & Sons, Inc.

SPONSORS

We gratefully acknowledge the support of:

Amgen
Bristol-Myers Squibb Company
Cytel Software Corporation
GlaxoSmithKline
ICON Clinical Research
Inspire Pharmaceuticals, Inc.
J & J PRD
Merck Research Laboratories
Millenium Pharmaceuticals
Novartis Pharmaceuticals, Inc.
Pfizer, Inc.
PPD, Inc.
Rho, Inc.
SAS Institute
Schering-Plough Research Institute
Statistics Collaborative, Inc.
The Emmes Corporation
John Wiley & Sons, Inc.

ENAR

Officers and Committees

January – December 2005

EXECUTIVE COMMITTEE -- OFFICERS

President	Peter Imrey
Past President	Marie Davidian
President-Elect	Jane Pendergast
Secretary (2005-2006)	Lance Waller
Treasurer (2004-2005)	Joanna Shih

REGIONAL COMMITTEE (RECOM)

President (Chair) Peter Imrey
Six ordinary members (elected to 3-year terms): + Stacy Lindborg (RAB Chair)

2003-2005	2004-2006	2005-2007
Stephen George	Bruce Craig	Gregory Campbell
Roderick Little	Amita Manatunga	Naisyin Wang

REGIONAL MEMBERS OF THE COUNCIL OF THE INTERNATIONAL BIOMETRIC SOCIETY

Marie Davidian, Walter W. Piegorsch, Ron Brookmeyer, Louise Ryan & Janet Wittes

APPOINTED MEMBERS OF REGIONAL ADVISORY BOARD (3-year terms)
Chair: Stacy Lindborg

2003 – 2005	2004-2006	2005-2007
Scarlett Bellamy	Hongshik Ahn	Barbara Bailey
Christopher R. Bilder	Brent Coull	Sudipto Banerjee
DuBois Bowman	Debashis Ghosh	Jason Connor
James J. Chen	Amy Herring	Todd Durham
Michael Daniels	Tom Loughin	Kirk Easley
Francesca Dominici	Jared Lunceford	Abie Ekangaki
Montserrat Fuentes	Jeffrey Morris	Deborah Ingram
Cynthia Garvan	Kerrie Nelson	Xuejen Peng
Brian D. Marx	Frank Roesch	James Rosenberger
Alicia Y. Toledano	Helen Zhang	Maura Stokes

ENAR

Programs

2005 Spring Meeting – Austin, TX
Program Chair: A. John Bailer
Program Co-Chair: Maura Stokes

2005 Joint Statistical Meeting
Naisyin Wang

2006 Spring Meeting – New Orleans, LA
Program Chair: Montserrat Fuentes
Program Co-Chair: José Pinheiro

2006 Joint Statistical Meeting
TBD

Biometrics Editors
Laurence Freedman, Mike Kenward, and Xihong Lin

Biometric Bulletin Editor
Urania Dafni

ENAR Correspondent for the Biometric Bulletin
Roslyn Stone

ENAR Executive Director
Kathy Hoskins

International Biometric Society Executive Director
Claire Shanley

Representatives

Committee of Presidents of Statistical Societies (COPSS)
ENAR Representatives
Peter Imrey (President) Marie Davidian (Past-President) Jane Pendergast (President-Elect)

ENAR Standing/Continuing Committee Chairs

Nominating	Marie Davidian
Sponsorship	Frank Shen
Information Technology Oversight (ITOC)	Bonnie LaFleur

American Association For the Advancement of Science (Joint with WNAR) Terms through February 22, 2008

Section E, Geology and Geography	Stephen Rathbun
Section G, Biological Sciences	Geof Givens
Section N, Medical Sciences	Joan Hilton
Section O, Agriculture	Kenneth Portier
Section U, Statistics	Mary Foulkes

National Institute of Statistical Sciences (ENAR President is also an ex-officio member) Board of Trustees
Member: Peter Imrey

Fostering Diversity Workshop

Co-Chair: Scarlett Bellamy
Co-Chair: Mahlett Tadesse
DuBois Bowman
Marie Davidian
Joel Greenhouse
Jacqueline Hughes-Oliver
Stacy Lindborg
Amita Manatunga
Dionne Price
DeJuran Richardson
Louise Ryan
Kimberly Sellers
Keith A. Soper
Tom Ten Have
Lance Waller

ENAR Student Award Committee

Chair: Timothy G. Gregoire
Karen Bandeen-Roche
Karl Broman
Murray Clayton
Brent Coull
Philip Dixon
Montserrat Fuentes
Susan Halabi
Daniel Hall
Xihong Lin
Jean Opsomer
Dionne Price
Alicia Toledano
Naisyin Wang

Student Award Winners

Van Ryzin Award Winner

Adin-Christian Andrei, University of Michigan

Award Winners

Jamie Lynn Bigelow, University of North Carolina at Chapel Hill
Wei Chen, University of Michigan
Jing Cheng, University of Pennsylvania
Andrea Cook, Harvard School of Public Health
Jia Guo, University of Minnesota
Xiang Guo, North Carolina State University
Ying Guo, Emory University
Xianzheng Huang, North Carolina State University
Mingyao Li, University of Michigan
Dawei Liu, University of Michigan
Qi Long, University of Michigan
Tanzy Love, Iowa State University
Haolan Lu, University of Minnesota
Brian Reich, University of Minnesota
Ronglai Shen, University of Michigan
Dan Sheng, New York University School of Medicine
Jie Yang, University of Florida
Lu Zheng, Harvard School of Public Health
Xin Zhi, University of Minnesota

ENAR

Future Meetings of the International Biometric Society

2005 WNAR Annual Meeting
June 21-24
Fairbanks, AK

2006 ENAR Spring Meetings
March 19-22
New Orleans, LA

2006 International Biometric Conference
July 16-21
Montreal, Quebec

2007 ENAR Spring Meetings
March 15-18
Miami, FL

2008 ENAR Spring Meetings
March 16-19
Crystal City, VA

2008 International Biometric Conference
July 14-19
Dublin, Ireland

SPECIAL THANKS

2005 ENAR Program Committee

A. John Bailer (Chair), Miami University
Maura Stokes (Co-Chair), SAS Institute
John Barnard, Cleveland Clinic Foundation
Dwight Brock, Westat
Cavell Brownie, North Carolina State University
James Chen, U.S. National Center for Toxicological Research
Donald Hedeker, University of Illinois at Chicago
J. Jack Lee, University of Texas MD Anderson Cancer Center
Stuart Lipsitz, University of South Carolina
Clarice Weinberg, U.S. National Institute of Environmental Health Sciences

ASA Section Representatives

Carmen Acuna, Bucknell University
Keaven Anderson, Merck & Company
Steve Anderson, U.S. Food and Drug Administration
Michael Daniels, University of Florida
Amita Manatunga, Emory University
Wendy Martinez, U.S. Office of Naval Research
Trivellore Raghunathan, University of Michigan
Glen Satten, U.S. Centers for Disease Control and Prevention
Machelle Wilson, University of Georgia

IMS Program Chair

Runze Li, Pennsylvania State University

ENAR Education Advisory Committee

A. John Bailer, Miami University
Alicia Carriquiry, Iowa State University
Allan Donner, University of Western Ontario
Robert Elston, Case Western Reserve University
Nancy Flournoy, University of Missouri
Nancy Geller, U.S. National Heart, Lung, and Blood Institute
Ramon Littell, University of Florida
Roderick Little, University of Michigan
Giovanni Parmigiani, Johns Hopkins University

ENAR Diversity Workshop

Scarlett Bellamy, University of Pennsylvania
Mahlet Tadesse, University of Pennsylvania

ENAR Student Awards

Timothy Gregoire, Yale University

Local Arrangements Assistance

Sarah Baraniuk
J. Jack Lee
Melissa Spann
Marina Vannucci

2005

ENAR Short Courses

SC1: Using Random Forests for Scientific Discovery
(Full Day: 8:30 am -5:00 pm) *Salon A*

Description:
Random Forests™ is arguably the most powerful current multipurpose tool for predicting and understanding data. A random forest is a collection of classification trees, each generated by simultaneous bootstrap sampling of a training set of available cases and repeated subsampling of a class of predictor variables. The approach originated about four years ago from research on ensemble methods in the Machine Learning community. It has gone through three years of development and is implemented in free open source (f77) software with extensive documentation (*//stat-www.berkeley.edu/users/breiman/RandomForests/*) and an interface to R.

Random Forests can be used for both classification and regression problems with thousands of variables with small or large sample sizes. Comparative tests of the Random Forests software (RF) show accuracy comparable to the best current prediction algorithms, such as Support Vector Machines. In addition to prediction accuracy Random Forests can detect and rank important variables, and offers other features that promote insightful data analysis.

For instance, derived intrinsic similarities between cases can be used to produce a two-dimensional data representation revealing unsuspected aspects of the data. These similarities are also helpful for imputing missing data and outlier detection. The user may also learn about which variables are driving the classification by examining the distribution of variables in high class density areas in this representation.

Data sets where the features of a rare class are of most interest are becoming more frequent. An innocent classifier will work on correctly classifying the uninteresting cases at the expense of a high error rate in the rare but important class. RF has an effective method for giving more balanced results in such highly unbalanced data.

A companion Java program supports powerful interactive graphics on output files from RF, letting the user conveniently explore local aspects of the data.

RF comes in separate versions for classification, nonlinear multiple regression, and model-free survival analysis. In the morning and early afternoon, the classification version of RF will be discussed. We will give an intuitive idea of why it works so well, overviews of the workings of the various methods and options mentioned above, and illustrations on real data sets including some comprehensive case studies.

DATE: Sunday, March 20, 2005

Full Day Fee	
Members	$200 ($225 after 2/20)
Nonmembers	$250 ($270 after 2/20)
Half Day Fee	
Members	$125 ($150 after 2/20)
Nonmembers	$165 ($190 after 2/20)

Short Course Registration
Saturday, March 19 3:00 - 5:00 p.m.
Sunday, March 20 7:00 - 8:30 a.m.

(lunch on your own)

This will be followed by a "How To Do It" session designed to familiarize the user with the software options, switches, data input procedures, and naming conventions, to the point where attendees can make informed use of the program.

The later afternoon is devoted to a discussion of the regression version of RF. Many aspects, such as variable importance, similarities etc. are common with the classification version but have different algorithmic implementations. An important capability is that of giving different weights to differing intervals of the response values. This allows an examination of which variables are important in varying ranges of the response. Again, we will give an overview of how things work followed by a "How To Do It" session.

Survival analysis is an important application of the regression program. The missing value method in RF is used to fill in censored times. The output pinpoints the important variables. We will close by illustrating the procedure, on a variety of data sets, both real and simulated.

Prerequisites: None except a desire to explore and understand data.

ENAR Short Courses

SC2: GLMM and GEE Modeling of Complex, Non-Gaussian, Correlated Data
(Full Day: 8:30 am -5:00 pm) *Salon G*

Instructors: José Pinheiro (Novartis Pharmaceuticals) and Edward Chao (Insightful Corporation)

Description:
Clustered, and hence correlated, data are routinely collected in an increasing number of areas of biostatistical application. Such data naturally arise from varied data collection methods including industrial laboratory experiments, multi-stage sample surveys, ecological and environmental surveys, observational health databases, clinical trials, community intervention studies, and genetic and other studies of families and households. When the response variable of interest can be assumed to follow a Gaussian distribution, linear and nonlinear mixed-effects models are generally used to fit and analyze such clustered data. However, these methods are not suitable for the non-Gaussian data that are the norm, rather than the exception, in many areas of investigation. Common examples of clustered non-Gaussian data include binomial, multinomial, and Poisson counts collected as repeated measures under different conditions, in multilevel sampling configurations, or in longitudinal follow-up.

Generalized Linear Mixed-effects Models (GLMMs) and Generalized Estimating Equations (GEE) models provide powerful tools for fitting and analyzing correlated non-Gaussian data, because they model flexibly the covariance patterns, including variance heterogeneity and hierarchical correlation structures, observed with such data. This course will provide an overview of the theory and application of GLMM and GEE models for the analysis of clustered non-Gaussian data, and compare these two approaches. A unified model-building strategy for the different types of models will be presented and applied to the analysis of a variety of real datasets including sociological data, panel data, spatial observations, and clinical trial outcomes.

The course will start with a brief review of generalized linear models for independent data and linear mixed-effects models for clustered Gaussian data. Single level and multilevel GLMMs will then be described, with different estimation methods (e.g., penalized quasi-likelihood - PQL, adaptive Gaussian quadrature - AGQ) discussed and contrasted. Bias and variability of the competing estimators will be compared through simulation results. The AGQ method and its efficient implementation will be explored in greater detail. Next, we will present GEE methods for multilevel data, variance components, and mixed-effects modeling. Applications to multivariate correlated data, such as multivariate binomial and multinomial longitudinal data and multimodal diagnostic testing methods, will be demonstrated. The GLMM and GEE approaches will then be illustrated and compared via simulation studies and parallel analyses of real data sets. Commercial software (e.g., S-PLUS and SAS) for modeling correlated non-Gaussian data will be presented, compared, and illustrated through examples. Advanced topics in mixed and GEE modeling, such as handling of missing data, smoothing methods, and fitting algorithms for very large datasets, will also be included to the extent time allows.

The presenters are the authors of the S+CorrelatedData library for GLMM and GEE models, currently available on an experimental basis pending incorporation into a future version of S-PLUS . Dr. Pinheiro is author, with D. Bates, of the text "Mixed-Effects Models in S and S-PLUS" on which some course material will be based.

Prerequisites: This will be an applied course, with emphasis placed on model-building methods and diagnostics rather than theory. Familiarity with linear models for independent data, including matrix notation, will be assumed. Some previous experience with generalized linear models for independent or uncorrelated data, and with linear mixed-effects models for correlated data, would be useful, but is not essential as these topics will be reviewed in the course.

SC3: Monitoring Clinical Trials
(Full Day: 8:00 am - 5:30 pm) *Salon J*

Instructor: Michael Proschan (National Heart, Lung, and Blood Institute, National Institutes of Health)

Modern comparative clinical trials are monitored for safety and efficacy as the data come in, rather than exclusively at the end. Monitoring typically includes formal interim statistical analysis. If such an analysis makes clear that one treatment is superior, or that no treatment difference will emerge by the end of the trial, considerations of ethics and resource allocation may require stopping the trial early and reporting its results. This course deals with how one decides whether evidence is sufficient to do that.

The course begins by showing that many clinical trial statistical tests, including t-tests, tests of proportions, some survival tests, and tests based on linear and mixed models, fall within a single, general Brownian motion framework. We will begin with a simple example, and then relax assumptions to generalize to other settings. The Brownian motion paradigm will be used to motivate classical boundaries (Haybittle-Peto, Pocock, O'Brien-Fleming) as well as the general spending function approach of Lan & DeMets, and to compute conditional power, an essential tool for determining whether trial continuation is futile. Many examples will be given to show the step-by-step process of using Brownian motion to compute conditional and unconditional power, and how these are used to decide whether a trial should stop early.

ENAR Short Courses

Other classical topics include methods of computing p-values, estimating parameters, and computing confidence intervals following a group-sequential trial. We will show these are problematic because the sufficient statistic is now a pair—the time of stopping and the z-score. There is no unique way to order sample outcomes, and different orderings lead to different inferences. Nor can group-sequential trials produce minimum variance parameter estimates.

Much of the afternoon will be devoted to discussion of less standard monitoring techniques including small sample methods like group-sequential permutation tests, some methods of adaptively modifying sample size in response to interim assessment of either nuisance parameters (e.g., variability, baseline hazard) or effect size, and Bayesian monitoring methods. (See also Tutorial T5 for a detailed focus on Bayesian monitoring.)

Note that, due to its breadth of coverage, this course will begin a half-hour earlier (8 a.m.) and end a half-hour later (5:30 p.m.) than SC1 & SC 2.

Prerequisites: The course is intended for statisticians familiar with tests commonly used in clinical trials, including t-tests, tests of proportions, and the logrank test. No prior knowledge of monitoring is assumed.

SC4: The Statistical Evaluation of Surrogate Endpoints in Clinical Trials
(Half Day: 8:00 am -12:00 noon) *Salon K*

Instructor: Geert Molenberghs (Limburgs Universitair)

Description:
Both humanitarian and commercial considerations have spurred intensive search for methods to reduce the time and cost required to develop new therapies. The identification and use of surrogate endpoints, i.e. measures that can replace or supplement other endpoints in evaluations of experimental treatments or other interventions, is a general strategy that has stimulated much enthusiasm. Surrogate endpoints are useful when they can be measured earlier, more conveniently, or more frequently than the "true" endpoints of primary interest (Ellenberg and Hamilton, Statistics in Medicine, 1989). Regulatory agencies around the globe, particularly in the United States, Europe, and Japan, are introducing provisions and policies relating to the use of surrogate endpoints in registration studies. But how can one establish the adequacy of a surrogate, in the sense that treatment effectiveness on the surrogate will accurately predict treatment effect on the intended, and more important, true outcome? What kind of evidence is needed, and what statistical methods portray that evidence most appropriately?

The validation of surrogate endpoints has been studied by Prentice (Statistics in Medicine, 1989), who presented a definition of validity as well as a formal set of criteria that are equivalent if both the surrogate and true endpoints are binary. Freedman, Graubard, and Schatzkin (Statistics in Medicine, 1992) supplemented these criteria with the proportion explained which, conceptually, is the fraction of the treatment effect mediated by the surrogate. Noting operational difficulties with the proportion explained, Buyse and Molenberghs (Biometrics, 1998) proposed instead to use jointly the within-treatment partial association of true and surrogate responses, and the treatment effect on the surrogate relative to that on the true outcome. In a multi-center setting, these quantities can be generalized to individual-level and trial-level measures of surrogacy. Buyse et al. (Biostatistics, 2000) have therefore proposed a meta-analytic framework to study surrogacy at both the trial and individual-patient levels.

This course will present an overview of these developments, with illustrations from the fields of ophthalmology, oncology, and mental health. The presenter is co-editor of a forthcoming reference on this topic (Burzykowski, Molenberghs, and Buyse, "The Evaluation of Surrogate Endpoints", New York: Springer-Verlag, anticipated publication first quarter 2005 or shortly thereafter) and co-author on a number of chapters in the text.

Prerequisites: Familiarity with standard statistical concepts, such as hypothesis testing, parameter estimation, and linear and logistic regression. Some knowledge of hierarchical models (such as linear mixed models) will be helpful, but not strictly necessary.

SC5: Up-and-Down Procedures and Some Other Response-Adaptive Designs
(Half Day: 1:00 - 5:00 pm) *Salon B*

Instructor: Nancy Flournoy (University of Missouri)

Description:
This course introduces the theory and application of up-and-down procedures and some other response-adaptive designs for use in laboratory and human experimentation. In these experiments, subjects arrive sequentially or in groups. Typically response to treatment is observed before arrival of the next individual or group, and treatment assignments depend on prior outcome(s). However, some recent results covering delayed responses will also be discussed. In toxicology experiments and Phase I and II clinical trials, the dose of an agent can be varied and may itself be an object of investigation. In those settings we focus on up-and-down designs in which treatment assignments are never more than one dose level distant from the dose used most recently. For Phase III clinical trials, we present several response-adaptive procedures that supercede the Randomized Play-the-Winner-Rule.

For all procedures presented, we show how to control the treatment distribution realized by the adaptive allocation design, and reconcile considerations of efficiency and ethical constraints. We also discuss procedures and issues regarding estimation and inference once the data are in.

It will first be assumed that the probability of toxicity increases with dose, as is typical in Phase I clinical trials and toxicology studies, and that subjects arrive individually. In these experiments the goal is to estimate the maximally tolerated dose (MTD), or alternatively the dose that can be expected to produce a defined toxic response in 100p% of subjects (LD100p or TD100p), where p is a prespecified toxicity rate. For laboratory studies, the target dose is usually the LD50. Extensions, including applications to studies where subjects arrive in groups and where non-Markov assignment rules are employed, will also be presented.

When the probability of success (treatment efficacy without toxicity) is unimodal with treatment dose, as is often reasonable to assume for Phase II clinical trials under a continuation ratio or contingent response model, we will show how to design rules that cluster the realized treatment assignments around the dose that maximizes the chance of success. Finally, post Randomized Play-the-Winner procedures for two-arm clinical trials will also be presented and discussed.

Prerequisites: Basic knowledge of probability and statistics at the first year graduate level (e.g., Hogg and Craig, Casella and Berger, Rice, etc.). All one needs to know about stochastic processes in order to understand the behavior of these procedures will be simply described. Applications will be stressed.

SC6: Statistical Analysis of DNA Sequences
(Half Day: 1:00 - 5:00 pm) *Salon K*

Instructor: Spencer Muse (North Carolina State University)

Description:
The past decade has seen an explosion in the amount of DNA sequence data available to biologists. The availability of complete genome sequences of organisms including human, mouse, fruit fly, and the model plant Arabidopsis thaliana has opened up many avenues of biological research. The simultaneous influx of data and expansion of research goals have led to an increased need for sound statistical analyses, while the long-term rapid increase in computing power per unit cost has continued apace. These circumstances have resulted in a fertile field of investigation, with the proliferation of data and open research questions outpacing the number of individuals qualified to carry out statistical research.

This course will be organized around the key biological questions facing genome scientists. After a brief introduction of the necessary biological background, current methods for addressing each topic (including major bioinformatics software packages, e.g., BLAST, hmmer, clustalw) will be surveyed. Statistical approaches will be emphasized, although more purely computational areas will also be covered. Open questions and research opportunities will be identified along the way.

The course will focus on three primary questions:

1. How do we "annotate" the functions of DNA regions in genomes? We will discuss methods for finding genes in very long DNA sequences, as well as methods for predicting the functions or physical structures of those genes once they have been found. Applications of hidden Markov models will be highlighted.
2. How do we determine if two or more DNA sequences are biologically similar? This topic includes the important problem of sequence alignment (and its close relative, database searching). Computational algorithms will be presented, and the statistical properties of scores arising from those algorithms will be discussed.
3. How do we analyze data from the genomes of multiple organisms simultaneously? This body of methods is known as "comparative genomics," and is gradually becoming more and more statistical. We will survey a variety of applications, but focus attention on the problems of estimating evolutionary trees and of using previously annotated genomes to assist in the annotation of newly sequenced genomes.

The presenter is co-author (with Greg Gibson) of "A Primer of Genome Science" (Sinauer Associates Inc., 2002), the second edition of which will appear prior to the course, and on which some course material will be based.

Prerequisites: Participants should have an understanding of MS level mathematical statistics. A basic understanding of molecular genetics will be helpful but not necessary. Participants with no prior exposure to genetics might benefit by reading a very basic introduction, e.g., Gonick and Wheelis' "The Cartoon Guide to Genetics" (Harper Collins, 1991), before attending.

ENAR 2005 Tutorials

T1: Quantile Regression (Monday, 8:30 - 10:15 am)

Salon F

Instructors: Xuming He (University of Illinois at Urbana-Champaign) and Ying Wei (Columbia University)

Description:
Regression analyses examine how distributions of a response variable change with conditioning on values of one or more predictors. Linear regression and other generalized linear models attempt to portray such changes solely through their effects on the conditional mean or through specification of a global parametric likelihood, while global survival models such as the proportional hazards model portray average effects over an entire population. In contrast, quantile regression portrays the influence of predictors on the conditional quantiles, such as the median or the 90th percentile. In many situations where information on the tails of a distribution is more important than its center, for instance in environmental toxicology and clinical medicine, quantile regression methods may directly yield important insights unavailable or elusive when analytic methods directed at population means are used.

Quantile regression is particularly useful for modeling and analyzing the relationship between a response variable and its covariates in heterogeneous populations, where the upper and lower tails of the conditional quantile functions may behave very differently from the central trend. For example, a treatment may be quite advantageous for individuals with longer survival times but ineffective for those with shorter times. Such differences are easily missed by models where effects are averaged over the whole population. Quantile regression makes minimal assumptions on the likelihood function, and thus offers the flexibility to capture heterogeneity, bimodality, and other potentially important features that are of interest in the agricultural, biological, environmental, and health sciences.

This tutorial will begin with motivating examples for quantile regression, followed by a careful explanation of the concepts. Using widely available R and SAS software, we will show how to use quantile regression in linear, nonlinear, and possibly nonparametric models. Examples including a recent application to longitudinal growth charts will be discussed. The common problem of data censoring will also be addressed using accelerated failure time (AFT) models.

Prerequisites:
A working knowledge of linear regression analysis is required. Familiarity with any of SAS, R, or S is a plus.

T2: Sample-Size Analysis in Study Planning with SAS PROCs POWER and GLMPOWER: Concepts and Issues, with In-Depth Examples (Monday, 1:45 - 3:30 pm)

Salon F

Instructors: Ralph G. O'Brien (Cleveland Clinic Foundation) and John M. Castelloe (SAS Institute)

Description:
Prospective sample-size analysis is invaluable to research design, promoting wiser allocation of scientific resources and stronger bioethics. Moreover, the process itself induces excellence and breadth in scientific planning by requiring the research team to delineate, critique, and tighten the research questions, study rationale, and many aspects of study design, including outcome measurements and analysis plans. Critically, the team must make reasonable conjectures about the "infinite datasets" representing the study populations. This is supported by ever-improving methods and software for computing power and/or required sample sizes under multiple plausible scenarios. In contrast, ritualistic sample-size and power computations, promulgated with little consideration of scientific context, are an empty exercise at best.

This tutorial will demonstrate the use of PROCs POWER and GLMPOWER (new in SAS 9), augmented with other SAS modules/macros developed and distributed free by the presenters. In-depth examples will illustrate how to use this software, all within the context of the science at hand. This involves, in part, (1) positioning a study in a line of scientific investigation (early to middle to late in "The March of Science"); (2) sizing a study for precision of an statistical interval or power of a conventional hypothesis test; (3) considering positive and negative inference mistake rates (false discovery and false miss rates) rather than traditional Type I and II error rates; and (4) communicating power and sample size concepts and results to non-statistician investigators. Non-SAS users should have no difficulty applying these notions to other software systems.

Although the methods covered here will be frequentist, basic Bayesian concepts help to clarify and shape the planning process for both investigators and statisticians. Those wishing to learn how to perform sample-size analyses completely from a Bayesian perspective should also consider Tutorial T5, which has been planned to complement this session.

Prerequisites:
Familiarity with hypothesis testing and confidence intervals. Prior exposure to SAS will help, but is not necessary. The tutorial is intended for statisticians seeking tools to support and increase the relevance of power and sample size analyses, and better integrate them into study planning through improved communication with investigators.

ENAR 2005 Tutorials

T3: Tools and Tips for Handling Good Data With Bad Properties: Analytic Approaches for Health Care Cost, Occupational/ Environmental Exposure, and Biomarker Data in the Real World (Monday, 3:45 - 5:30 pm)

Salon F

Instructor: Xiao-Hua Andrew Zhou (University of Washington and Seattle Veterans Administration Medical Center)

Description:
This tutorial will briefly survey some statistical approaches for handling several commonly encountered features of data that may preclude, or at least compromise the validity, of standard analyses. Such features include skewed distributions, heteroscedasticity, clumping of data at one particular value (e.g., zero), and informative censoring. For example, health care costs, mineral resources in the earth's crust, air pollutants, species abundances, and toxic occupational exposures often exhibit skewed distributions, non-constant variances, or both. For variables such as health care costs, occupational exposures, and chemicals or biomarkers measured by analytic techniques with lower detection limits, nontrivial portions of a study population may be expected to have values of zero, or effectively zero, because the subjects do not get sick, are not exposed, or the exposure is present in an undetectable concentration. Finally, an outcome may be informatively censored, e.g., total health care costs may be censored by the termination of a study observation period. Such censoring is informative because subjects hospitalized for extended durations incur very high costs and are most likely to be censored. Naive statistical analyses of data with these characteristics are prone to erroneous inferences and predictions. We will discuss statistical issues underlying such problems and introduce recently developed methods for handling these distributional features. In particular, we will consider:

- Point and interval estimation for a skewed population with or without censoring;
- Interval estimation for one skewed population with additional zero values using transformation methods;
- Interval estimation and testing, including multiple comparisons, for two or more skewed populations with and without additional zero values;
- Parametric and semiparametric regression models for skewed heteroscedastic populations with and without additional zero values;
- Parametric and semiparametric regression models for skewed populations with censoring.

The presentation will be oriented to the data analyst, with examples.

Prerequisites:
A one year course on mathematical statistics, and familiarity with likelihood- based inference.

T4: Statistical Methods For Gel Electrophoresis Proteomics (Tuesday, 8:30 - 10:15 am)

Salon F

Instructor: Francoise Seillier-Moiseiwitsch (Georgetown University)

Description:
The term "proteome" has been coined to reflect the revolutionary changes the field of biochemistry has been undergoing. This word refers to the PROTEins expressed by a genOME or tissue. Unlike the genome, the proteome is greatly affected by numerous factors such as tissue and environmental conditions. A gene can be spliced in many different ways and proteins can be altered after translation (i.e., gene-directed synthesis). Hence, the proteome consists of far more proteins than the genome contains genes. Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), and Surface Enhanced (SELDI) and Matrix Assisted (MALDI) Laser Desorption/Ionization mass spectrometry, are currently the techniques of choice to separate and display all the proteins expressed in a tissue.

Separately, invited paper sessions 14 and 25 will provide an introduction to mass spectrometry proteomic data analysis. This tutorial will describe statistical issues relating to the underlying technology and analysis of 2D-PAGE data. In 2D-PAGE, proteins appear as spots on an image, separated on one dimension by molecular weight, and on a second dimension by isoelectric point, that is, the pH value at which the protein exhibits no net charge and is least soluble. In the resulting protein map, individual protein spots are identified and quantified.

We will introduce the technology involved in two-dimensional gel electrophoresis, including the specific steps in data acquisition. Features of the raw data, and themes common to the two areas of proteomics, 2D-PAGE and mass spectrometry, will then be discussed. Review of image analysis procedures for image segmentation, registration, and normalization will be followed by an overview of the statistical and computational techniques used to identify differentially expressed proteins. The limitations of each technique will be highlighted using analyses of published data sets. Similarities and differences with the analysis of microarray data will be noted.

The instructor is co-author (with Anindya Roy) of the forthcoming book, "Statistical Methods for Proteomics", to appear in Summer 2005 as a volume in the SIAM-ASA Series in Statistics and Applied Probability.

Prerequisites:
Basic statistical background and an interest in understanding the basic biology, statistical pitfalls, and data analytic tools relevant to proteomic data.

ENAR 2005 Tutorials

T5: Bayesian Approaches for Clinical Trial Design and Analysis (Tuesday, 1:45 - 3:30 pm)

Salon F

Instructor: Bradley P. Carlin (University of Minnesota)

Description:
Thanks in large part to the rapid development of Markov Chain Monte Carlo (MCMC) methods and software for their implementation, Bayesian methods have become ubiquitous in modern biostatistical analysis. In clinical research, Bayesian methods have been in common use for over a decade in submissions to the FDA Center for Devices and Radiological Health, where data on new devices are often scanty but researchers typically have access to large historical databases. Statisticians and regulators on the drug side of FDA are also coming to appreciate the utility of these methods for combining information from separate but related sources, reducing necessary sample size, and directly measuring effects of interest while protecting overall error rates.

In this tutorial we will review the features associated with Bayesian methods and give specific examples of the potential benefits they offer in drug and device trials. In particular, we will show how Bayesians think about "power" when designing a trial, and how a Bayesian procedure may be calibrated to guarantee good long-run frequentist performance (i.e., low Type I and II error rates), a subject of keen interest to the FDA. We will also discuss common methods for specifying prior distributions, as well as show how Bayesians do interim monitoring (thus answering the time-honored frequentist question, "Why do Bayesians get to peek at their data and I don't?"). Finally, we will indicate a new way of combining the two most popular pieces of software used by the Bayesian community, R and WinBUGS, in order to handle more complex model settings.

Prerequisites:
Tutorial participants should have an M.S. understanding of mathematical statistics at, say, the Hogg and Craig (2000) level, as well as basic familiarity with standard statistical models, computing, and traditional approaches to clinical trial monitoring and power analysis. Those with knowledge of Bayesian software and methods (say, based on the WinBUGS software and the book by Carlin and Louis, 2000) will face a gentler learning curve. However, no extensive previous exposure to Bayesian methodology will be assumed. The material should be accessible to those familiar with elementary concepts such as Bayes' rule, conjugate prior distributions, prior to posterior updating, the beta-binomial distribution, and elementary Bayesian calculations.

ENAR 2005 Roundtables

Date: Monday, March 21, 2005
Time: 12:15 - 1:30 pm *Salon H*

R1: Technology in Teaching Statistics and Epidemiology to Health Professionals: Performance and Prospects

Discussion Leader: David G. Kleinbaum, Emory University

Possibly no other field of science has been affected more directly than Statistics by the computing and communications revolutions, and every aspect of epidemiologic method and practice has also been dramatically altered. Most would agree that training in these fields has also been enhanced by the use of computers i) for instructional analyses and exercises, ii) to enhance didactic presentations, iii) to access Internet-based resources, and iv) to allow more frequent and rapid communications among students and between students and instructors. However, the degree to which the biostatistical and epidemiological capabilities of health care professionals have been affected is unclear, both in an absolute sense and relative to the investment of instructor time in technological improvements. Certainly use of the term "revolution" in this context is premature. This roundtable group will discuss the implications of technological change for future training of health professionals, including health scientists, in Biostatistics and Epidemiology. What is known about the efficacy of computer-based interactive instructional software in biostatistical and epidemiological training? What is the best role of data analytic and/or instructional software in service courses? Do CD ROM instructional texts such as ActivEpi and ActivStat and/or web-based instructional tools such as Blackboard, WebBoard, and others enhance learning in on-site courses? Should we embrace or shun distance education via computer for teaching these subjects to health professionals? To what extent do we have the power to shape the future instructional environment in these fields? What further innovations should we look for or attempt to stimulate? David Kleinbaum, the discussion leader, is well-known for innovative textbooks and teaching on epidemiological methods, multiple linear and logistic regression, and survival analysis, and for extensive worldwide short-course training in epidemiological methods. He is developer of the ActivEpi interactive computer-based instructional software (2002), which has been used in a variety of educational environments including distance learning.

R2: Reproducibility of Gene Expression in Microarray Studies.

Discussion Leader: Leslie Cope, Johns Hopkins University

In recent years, numerous gene expression microarray studies have associated many different genes with various diseases or other health-related phenotypes. However, different studies of the same disease phenotype often find different or even non-overlapping sets of associated genes, calling into question the validity of results from the disparate studies. Such discrepancies can arise for three reasons. Phenotypes may differ across studies because, inadvertently, the samples analyzed have differed in composition. Or, the studies may be using different microarrays with few genes in common. Finally, the genes may be inaccurately measured in at least one study. This discussion will focus on this latter situation. How can we assess whether or not genes are reliably (i.e., reproducibly) measured in gene expression experiments, and thereby identify such measurement errors? How can we assess the genesis of measurement error when it has been identified? What methods are available for prevention or correction of such errors? Some recently published methods for using several studies to determine gene reproducibility will be described and compared, and microarray reproducibility further discussed depending on the interests of participants.

R3: Biostatistics and Biometrics for Biosurveillance and Biosecurity Applications

Discussion Leader: G.P. Patil, Pennsylvania State University

The 21st Century has brought unprecedented terrorist threats to the physical security of developed nations, the United States especially, through possible local delivery of nuclear, chemical, and/or biological weapons. Response to these threats requires improved systems to identify, locate, and intercept terrorists and their activities and, as a second line of defense, to recognize and attenuate attacks as or shortly after they occur. Various "biometric" identification tools, such as routine fingerprint entry systems, dynamic signature recording, and iris scanning, have been proposed for the former purpose. For recognizing biological attacks on infrastructure and homeland security, increased and more effective defense may require dynamic geospatial and network surveillance of crop pathogens and invasive species, human infectious disease organisms, and radiation, through distributed sensor and biosensor networks and the public health system. This roundtable will discuss the technical statistical challenges, and opportunities for the biostatistical/biometric community to contribute through research, education, and outreach, presented by biometric identification, biosurveillance, and other activities associated with homeland security efforts. The discussion leader is Director of the Penn State University Center for Statistical Ecology and Environmental Statistics, Editor-in-Chief of Environmental and Ecological Statistics, and an active researcher in geographic and network surveillance using scan statistics for the identification and prioritization of geospatial "hot spots." Those interested in this roundtable may also be interested in the Invited Paper Sessions, "Biosurveillance GeoInformatics for Biosecurity" and "At the Crossroads of Biostatistics and Security Biometrics."

ENAR 2005 Roundtables

R4: Analysis of Longitudinal Data in Clinical Trials with High Mortality

Discussion Leader: Tom Greene, Cleveland Clinic Foundation

Analysis of change in continuous longitudinal outcomes is often complicated by early termination of follow-up due to death. Informative-censoring models may be used to compare mean changes between treatment groups while accounting for the relationship between death and the outcome variable. A key drawback, however, is that such models evaluate mean changes that would have occurred in mortality-free, hence implausible, virtual populations. Alternatively, mean changes to a designated time may be evaluated conditionally on survival. However, the conditional approach violates the intent-to-treat principle because the analyses are restricted to survivors. This roundtable will discuss methods that attempt to avoid these difficulties. Possible approaches include a) recasting the problem as a time-to-event analysis of a composite endpoint that incorporates both the longitudinal outcome and mortality; b) evaluating parameters, such as slopes, that characterize trajectories rather than mean changes; c) jointly analyzing the longitudinal and mortality data without inference to marginal means; and d) using the Rubin-Causal model to estimate mean changes in patients who "would have survived" under each intervention. The discussion will reveal advantages and limitations of these approaches through examples from recent randomized trials.

R5: What Can We Learn from Heterogeneity of Effects Across Studies in Meta-Analyses?

Discussion Leader: Jesse A. Berlin, Johnson and Johnson Pharmaceutical Research and Development, LLC

Above-random heterogeneity is ubiquitous among findings of similarly-targeted studies. Thus, when undertaking meta-analysis, one should probably assume that quantifying heterogeneity and identifying its possible sources will be an important part of analysis. In this context, the distinction between heterogeneity due to modifiable aspects of study conduct and analysis, and that due to real variation in treatment effect, is critical. The former would include variation due to design issues such as how blinding is implemented, and analytic issues such as whether relative risk or risk difference is the more appropriate scale for treatment effect. Variation due to clinical factors, in contrast, is generally intrinsic to the underlying biology and, when recognized, offers potential to target therapy to the most appropriate patients. We'll discuss the discovery of unexpected variation across studies and, particularly, some specific meta-analyses that have revealed heterogeneity from both design features and treatment effects, and where understanding how these relate has informed policy or practice.

R6: Hierarchical Models in Health Services and Outcomes Research (HSR): The Unresolved Issues

Discussion leader: Constantine Gatsonis, Brown University

Hierarchical Modeling (HM) has become a standard methodology in Health Services and Outcomes Research (HSR). The use of hierarchical models, motivated by the natural multilevel structure of HSR data, has enabled investigators to combine data across health care providers, to account for clustering, and to examine the effects of characteristics of units, such as patients and hospitals. The underpinning of the more important contributions to the methodologic literature on measuring and comparing quality of care across health care providers arguably rests on ideas and methods from hierarchical modeling. However, despite the centrality of HM to the scientific method in this area of research, reports on quality of care - such as those from government agencies and companies producing "report cards" - are based on simpler (and often simplistic) methods of dubious scientific validity. Indeed, more generally it appears that the closer one gets to determination and implementation of actual health care policy and reporting of quality of care data to the public, the less often one sees the use and influence of hierarchical modeling methods. This roundtable will attempt to identify factors that contribute to this apparent misalignment. Undoubtedly, some of these factors are connected with limitations of the theoretical models and computational techniques. Broader factors relating to speed of integration of statistical advances into interdisciplinary research, and subsequently into reports that inform policy and the public, also surely contribute. Working from the quality of care and other examples where its influence on policy has been limited, we will explore i) features of the hierarchical modeling approach that may hinder its acceptance in policy formulation, ii) aspects of statistical methods more generally that promote or hinder their acceptance by health services decision-makers, and iii) ways for statisticians to be most effective in bringing insights gained from modern data analytic tools to the forefront of health care policy formulation.

R7: How Many Statisticians Does it Take to Do a t-Test? - Conflicts of Statistical Roles in Industry-Sponsored Clinical Trials

Discussion Leader: Janet Wittes, Statistics Collaborative

This roundtable will consider the many statistical roles in industry-sponsored clinical trials. We will address conditions under which a company can perform all statistical tasks in-house. For studies in which some or many roles are outsourced, we will discuss how best to orchestrate everyone's roles and responsibilities, paying particular attention to clashes of cultures among the players. In the most complex cases, where industry sponsors the trial but an academic group runs it, serious conflicts and confusion may

arise with regard to regulatory responsibilities, ownership of data, publication policy, and writing of reports for the FDA and non-US regulatory authorities. On occasion, statisticians in these complex settings are asked to run other people's programs, merely tacking the treatment code to datasets and producing reports. Participants will identify the statistical roles that can arise within various organizational structures, and try to formulate conditions under which each is professionally acceptable.

R8: Regulations and Research: How Can the Two Coexist?

Discussion Leader: Carolyn Apperson-Hansen, Cleveland Clinic Foundation

Protected health information, de-identification, electronic records, electronic signatures.... These are the latest buzz words in the world of clinical research. What do they mean to you? The HIPAA Privacy and Security Rules as well as the FDA regulation, 21 CFR Part 11 create major challenges in the clinical research environment. For example, a recent report suggests that the HIPAA rule has the potential to decrease response rates to clinical follow-up surveys by at least 50%! So, how do we meet these challenges? We must understand the impact of the regulations on clinical research, balance a web of federal regulations, and show a good faith effort in becoming compliant. After a brief summary of the regulations, there will be open discussion of how best to manage these regulations while continuing to do clinical research, and an opportunity to share experiences with colleagues on what approaches to compliance have been most, and least, successful. The discussion leader presents training and information courses on 21 CFR Part 11 and HIPAA throughout the Cleveland Clinic Foundation and nationally. She is a member of a PDA (an international organization to advance pharmaceutical and biopharmaceutical technology) Task Group working with FDA advisors to implement 21 CFR Part 11 compliant GERM (Good Electronic Records Management practices) within FDA-regulated corporations.

R9: Probability vs. Center-Based Sampling in Longitudinal National Health Research

Discussion Leader: Rod Little, University of Michigan

The National Children's Study (*www.nationalchildrensstudy.gov*), currently in its planning stage, will examine the effects of environmental influences on the health and development of more than 100,000 children across the United States, following them from before birth until age 21. The study defines environment broadly, including natural and man-made environment factors, biological and chemical factors, physical surroundings, social factors, behavioral influences and outcomes, genetics, cultural and family influences and differences, and geographic locations. It will yield one of the richest information resources available for answering questions related to child health and development, and will form the basis of child health guidance, interventions, and policy for generations to come. In July of 2004 it was announced that locations and participants for the National Children's Study will be selected by national probability sampling. The question of whether this and similar studies should employ such probability sampling of the US population, rather than the "medical center" model that is more usual in such studies, is controversial and, for the National Children's Study, was widely debated. Why is this question so difficult to resolve? The roundtable discussion will elucidate the issues surrounding this debate, and likely similar debates pertaining to future studies.

R10: What Topics Should Be in a One-Time Genomics Course for Biostatisticians?

Discussion Leader: Michael Newton, University of Wisconsin

Research in biostatistical methods increasingly concerns the genomic sciences, and much collaborative activity demands a basic understanding of genome biology. Successful biostatistical training programs must include material from the genomic sciences, but the number of potential topics makes formulation of a "genomics for biostatisticians" course challenging and potentially controversial. We'll consider which issues and material to include, and how they might be organized. Candidate course topics come from three distinct areas: gene-mapping studies; the measurement, analysis, and synthesis of sequence data; and gene expression and proteomics. It is difficult to select among classical elements of statistical genetics for gene-mapping studies (linkage and association, fine-mapping, haplotyping, family-based methods, population-based methods, and methods for controlled crosses), because all are relevant to modern studies considering genome-wide data and high resolution SNP markers. Among methods for sequence data, many are considered part of "bioinformatics," so overlaps with computer-science training approaches must be considered. Moreover, rather than structure a course around data types and statistical methods used, one might choose to organize around either i) a framework formed by biological issues and how data are used to address them (e.g., what tools are needed to establish a transcription factor's role in a particular tissue?), or ii) statistical issues in construction and use of genomic resources, for example, in the development and manipulation of genomic data archives. Such choices give much food for thought, and discussion!

ENAR 2005 Roundtables

R11: The Canadian Experience With Accreditation of Professional Statisticians

Discussion Leader: Judy-Anne Chapman, Universities of Toronto and Waterloo

This roundtable will consider the Statistical Society of Canada's (SSC) experience with accrediting individual professional statisticians at the one-year anniversary of that program, and the extent to which the SSC program and experience are relevant to US professionals. Since 1993 some form of certification or accreditation of individual professional statisticians has been adopted by the Royal Statistical Society in the UK (at merger of the Society with the Institute of Statisticians), the Statistical Society of Australia, the Association des Statisticiennes et Statisticiens du Québec, and the SSC. US initiatives along such lines, most recently in 1993-94 within the American Statistical Association (ASA), have not succeeded. The issues affecting our discipline, and particularly the future outlook for professional statisticians, would seem similar on both sides of the US-Canadian border. ASA has thus recently established a Task Force to reexamine the certification/accreditation issue. The SSC offers two levels of accreditation: Professional Statistician (P.Stat.) and Associate Statistician (A.Stat.). These qualifications are intended to convey that the holder has achieved a certain level of professional competence in the understanding and application of statistical methods, and maintains a high level of ethical practice. The SSC website (www.ssc.ca) and two brochures advertise the existence of accreditation and the SSC Code of Ethical Statistical Practice to prospective applicants and employers. The rationale, requirements, and history of the SSC program, and its reception by the Canadian statistical and other professional communities during its first year, will be discussed with an eye to Canada's southern neighbor. The discussion leader, a biostatistician affiliated with the Universities of Toronto and Waterloo, is Chair of the Initial SSC Accreditation Committee.

R12: Whither Privacy, Confidentiality, Data Sharing?

Discussion Leader: Stephen E. Fienberg, Carnegie Mellon University

Statisticians working with medical records have always needed to deal with concerns regarding the privacy of patients whose data we analyze. During the last decade, public policy has shifted towards greater data sharing and access, e.g., through mandates for data archiving and release by NSF, NIH, and other research funding agencies, but there has been a reversal towards heightened vigilance in privacy protection in part as a consequence of the Health Insurance Portability and Accountability Act of 1996 (HIPAA). Beyond the specific issues of how researchers can cope with such a regulatory environment (to be discussed at Roundtable 8), what should statisticians be doing to influence and manage societal trade-offs between confidentiality and access? The National Institutes of Statistical Sciences (NISS) NSF-funded digital government projects embody some technical responses to this question. How far can such technical responses take us? Data archives provide one form of access and government statistical agencies have created restricted research data centers for related purposes. Are there principles and practices going beyond these technical responses, that statisticians should be articulating and working towards? Besides the individual obligation not to reveal personal data, do statisticians have ethical responsibilities in the formulation of confidentiality and access control policies? Come share your experiences. Learn how others have dealt with the apparent conflict between data access and sharing and privacy protection, and the associated costs. The discussion leader is a senior investigator in the NISS Digital Government projects on confidentiality and disclosure limitation and has written on data sharing, and technical aspects of disclosure limitation methods. He is also chair of the International Statistical Institute's Committee on Professional Ethics.

R13: The Biostatistician's Perspective on Ethical Practices in Collaborative Research

Discussion Leader: Carol Redmond, University of Pittsburgh

This roundtable focuses on ethical problems of particular importance to biostatisticians. Collaborating biostatisticians are likely, at some time, to encounter situations where ethical issues must be addressed; about half of respondents to a 1998 survey of biostatisticians claimed direct knowledge of fraudulent or deceptive practices within their work environments during the previous ten years (Buyse et al., Statistics in Medicine, 1999; Ranstam et al., Controlled Clinical Trials, 2000). Problems with epidemiologic studies and clinical trials were reported similarly often. Biostatisticians also face ethical dilemmas about practices that are questionable without meeting a strict definition of research misconduct. Such dilemmas may arise at any study phase, from conceptualization of design to reporting the work, and may involve clinical as well as biostatistical issues. Although professional societies including the American Statistical Association provide ethical guidelines for professional conduct, these are not sufficiently specific for many problems biostatisticians encounter in collaborative biomedical research. Some notable cases, such as that of Roger Poisson in National Surgical Adjuvant Breast Project clinical trials, will be briefly summarized to highlight complexities and potential impacts of research misconduct. Questions for discussion include: 1) What are the nature and magnitude of ethical concerns commonly faced by collaborative biostatisticians? 2) Do our expertise in research methodology, and role in data quality assurance, provide unique insights into ethical problems,

and how to prevent, identify, and resolve them? 3) Are there ways, as individuals or through professional societies, that we should more actively support biostatisticians dealing with difficult ethical issues?

R14: Machine Learning Techniques in Modern Biostatistical Application

Discussion Leader: Trevor Hastie, Stanford University

"Machine learning" refers to processes by which computers can locate structure in data, through an iterative process that improves as data accumulate. Typically the term is reserved for processes that "learn" about the nature of structure, beyond just increasing the precision of a fixed set of parameter estimates. Machine learning techniques have been studied in the overlapping literatures of statistics, computer science, psychological and neurophysiological modeling, and adaptive control theory. In statistics, the term currently refers to a collection of modern classification and regression tools including perceptron learning, classification and regression trees (CART), near-neighbor methods, random forests, multivariate adaptive regression splines (MARS), generalized additive models, bagging, boosting, lasso, least angle regression, neural networks, and support vector machines. Such methods often forego global optimizations under strong probabilistic and structural assumptions in favor of algorithmically-based tools, using local optimizations and intensive computing, that accommodate flexibly large numbers of variables with unknown nonlinear relationships. This roundtable will consider the current and future roles of machine learning techniques in biostatistical practice. What advantages do such techniques offer beyond classical model-fitting methods? What does theory tell us about why algorithmic techniques perform so well in complex prediction problems? When are classical methods superior? What areas of biostatistical application could benefit from more rapid introduction of machine learning tools? Are such tools apt to be as useful in confirmatory as in exploratory studies? The interests of those who attend in these and other issues will largely frame the discussion.

ENAR

Fostering Diversity in Biostatistics

Sunday, March 20, 2005, 12:30 - 5:00pm
Meeting Room 410

On Sunday, March 20, 2005 the Eastern North Atlantic Region (ENAR) of the International Biometric Society will be hosting a workshop entitled "Fostering Diversity in Biostatistics". The workshop will provide a forum for discussion of important issues related to diversity and will be held at the Hilton Austin in Austin, TX. Themes of the workshop will include career and training opportunities in (bio)statistics as well as mentoring, recruiting, and retaining students in related graduate programs. The workshop will target persons from traditionally under-represented ethnic groups who are currently in or interested in learning more about biostatistics. Undergraduates and faculty from undergraduate institutions as well as Chairs and Directors of graduate training programs are especially encouraged to attend. In the past, in addition to faculty and students from undergraduate institutions, representatives from various graduate schools, industry, pharmaceutical companies and government agencies have attended the workshop enabling networking opportunities, opportunities to exchange ideas and opportunities to learn more about graduate and summer programs for persons interested in advanced training programs in biostatistics.

Registration is required for the workshop and lunch will be provided. Additionally, undergraduates and their faculty mentors who register and attend the workshop may be eligible to have conference registration fees for the larger ENAR meeting waived. To learn more, please contact: Scarlett Bellamy (*sbellamy@cceb.upenn.edu*) or Mahlet Tadesse (mtadesse@cceb.upenn.edu).

Workshop on NSF Funding Opportunities for Biometricians

Sunday, March 20, 2005, 5:30-7:00 pm
Xuming He
Program Director of Statistics, National Science Foundation
Salon B

Although the National Science Foundation (NSF) does not normally support bioscience research with disease-related goals, it does offer ample funding opportunities for biostatisticians. If you live on what you knew about the NSF a few years ago, you may be missing out on great funding opportunities. Through the Statistics Program and several interdisciplinary research programs, the NSF funds statistical research in biological sciences, including research activities not traditionally funded by other agencies. Dr. He will provide an overview of NSF programs most relevant to ENAR members. Continuing programs, new programs, the structure of and tips on NSF review processes, and recent awards, will all be considered. Programs to be discussed will include the Joint Division of Mathematical Sciences/Biological Sciences/National Institute of General Medical Sciences Initiative To Support Research In The Area Of Mathematical Biology; Focused Research Groups In The Mathematical Sciences; Genes and Genome Systems; Biocomplexity in the Environment; Mathematical Sciences Postdoctoral Research Fellowships; and others. Dr. He will also seek suggestions on how the NSF can better help fund research in Biostatistics.

5:30-5:50	General Funding Opportunities and Recent Special Programs
5:50-6:00	Questions, Discussion, Clarifications About Programs
6:00-6:20	NSF Review Processes for Disciplinary and Interdisciplinary Research Proposals
6:20-6:30	Examples of Recent Awards in Biostatistics
6:30-7:00	General Discussion and Questions, Including Feedback to NSF

Program Summary

SATURDAY, MARCH 19

3:00 p.m.–5:00 p.m. *Grand Ballroom Pre-function Area*	Conference Registration

SUNDAY, MARCH 20

7:30 a.m.–6:30 p.m. *Grand Ballroom Pre-function Area*	Conference Registration
8:00 a.m.–12:00 noon *Salon K*	Short Course SC4: The Statistical Evaluation of Surrogate Endpoints in Clinical Trials
8:00 a.m. – 5:30 p.m. *Salon J*	Short Course SC3: Monitoring Clinical Trials
8:30 a.m.–5:00 p.m. *Salon A* *Salon G*	Short Courses SC1: Using Random Forests for Scientific Discovery SC2: GLMM and GEE Modeling of Complex, Non-Gaussian, Correlated Data
12:30 p.m.–5:00 p.m. *Meeting Room 410*	Diversity Workshop
1:00 p.m.–5:00 p.m. *Salon B* *Salon K*	Short Courses SC5: Up-and-Down Procedures and Some Other Response-Adaptive Designs SC6: Statistical Analysis of DNA Sequences
3:00 p.m.– 6:00 p.m. *Grand Ballroom Pre-function Area*	Exhibits Open
4:00 p.m.–7:00 p.m. *Meeting Room 412*	ENAR Executive Committee (Closed)
4:30 p.m.–6:30 p.m. *Meeting Rooms 406 & 408*	Placement Service
5:30 p.m. – 7:00 p.m. *Salon B*	Workshop on NSF Funding Opportunities for Biometricians
8:00 p.m.–11:00 p.m. *Salon H*	Social Mixer and Poster Session

MONDAY, MARCH 21

7:30 a.m.–8:30 a.m. *Salon H*	Student Breakfast
7:30 a.m.–5:00 p.m. *Grand Ballroom Pre-function Area*	Conference Registration
9:30 a.m.–5:00 p.m. *Meeting Rooms 406 & 408*	Placement Service
8:30 a.m.–5:30 p.m. *Grand Ballroom Pre-function Area*	Exhibits Open
8:30 a.m.–10:15 a.m. *Salon F*	Tutorial T1: Quantile Regression

ENAR

Program Summary

8:30-10:15 a.m.	Scientific Program
Salon G	1. Recent Developments in Sequential Clinical Trials Methodology
Salon A	2. Cutting-Edge Risk Assessment Issues and Methods
Salon B	3. Analysis and Issues in Matched Family Aggregation Studies
Salon C	4. *Biometrics* Special Invited Paper Session
Salon D	5. Lack of Fit Tests for Model Misspecification with Applications to Longitudinal Data and Survival Data Analysis
Salon E	6. Contributed Papers: Computational Methods
Meeting Room 410	7. Contributed Papers: Nonparametric Methods in Longitudinal and Survival Analysis
Meeting Room 415	8. Contributed Papers: Survey Data Methods
Meeting Room 412	9. Contributed Papers: Environmental and Ecological Applications
Salon J	10. Contributed Papers: Survival Analysis I
Salon K	11. Contributed Papers: Analyzing High-Dimensional Genomic Data
10:15 a.m.–10:30 a.m.	Break
Grand Ballroom Pre-function Area	
10:30 a.m.–12:15 p.m.	Scientific Program
Salon A	12. Recent Advances in the Analysis of Recurrent Events Data
Salon B	13. Natural Resource Estimation for Small Areas (Theme Session)
Salon J	14. Dealing with MALDI-TOF/SELDI-TOF Proteomic Data: Experimental Design and Low-Level Processing
Salon D	15. At the Crossroads of Biostatistics and Security Biometrics (Theme Session)
Salon C	16. New Developments in the Analysis of High-Dimensional Data
Meeting Room 415	17. Contributed Papers: Health Services Research
Salon E	18. Contributed Papers: Modeling Methods in Epidemiology
Meeting Room 410	19. Contributed Papers: Environmental and Toxicological Applications
Salon G	20. Contributed Papers: Adaptive Methods and Designs
Meeting Room 412	21. Contributed Papers: Analysis of Correlated Data
Salon K	22. Contributed Papers: Linkage Analysis
12:15 p.m.–1:30 p.m.	Roundtable Luncheons (Registration Required)
Salon H	
12:30 p.m.–4:30 p.m.	Regional Advisory Board (RAB) Luncheon Meeting (By Invitation Only)
Meeting Room 400	
1:45 p.m.–3:30 p.m.	Tutorial
Salon F	T2: Sample-Size Analysis in Study Planning with SAS PROCs POWER and GLMPOWER: Concepts and Issues, with In-Depth Examples
	Scientific Program
Salon A	23. Statistical Issues and Novel Methods in Vaccine Clinical Trials
Salon K	24. Statistical Analysis of Wildfire Data (Theme Session)
Salon C	25. Bayesian Statistical Modeling of Mass Spectrometry Proteomic Data
Salon B	26. New Approaches to Statistical Access to Data in a Confidential World (Theme Session)
Salon D	27. Non-Normal Random Effects Models
Salon E	28. Contributed Papers: Methods in Epidemiology
Salon G	29. Contributed Papers: Clinical Trials I
Meeting Room 415	30. Contributed Papers: Spatial Modeling of Disease
Meeting Room 410	31. Contributed Papers: Recurrent Events Analysis
Meeting Room 412	32. Contributed Papers: Missing Data in Longitudinal Data Analysis
Salon J	33. Contributed Papers: Multiple Testing and False Discovery Rates
3:30 p.m.–3:45 p.m.	Break
Grand Ballroom Pre-function Area	

Program Summary

3:45 p.m.–5:30 p.m.	Tutorial
Salon F	T3: Tools and Tips for Handling Good Data with Bad Properties: Analytic Approaches for Health Care Cost, Occupational/Environmental Exposure, and Biomarker Data in the Real World
	Scientific Program
Salon A	34. Combining Spatial Measurement Parameters and Design of Experiments to Evaluate the Effectiveness of Treatments for Precision Agriculture (Theme Session)
Salon B	35. Spatial Epidemiology
Salon J	36. Statistical Concerns Under the Federal Advisory Committee Act (FACA)
Meeting Room 415	37. Statistical Analysis of Diffusion Tensor Imaging
Salon K	38. Recent Developments on Methods for Multivariate Failure Time Data
Salon D	39. Special Contributed Session: Advances in Applications of Latent Variable Model Analysis for Health Services Research
Meeting Room 412	40. Contributed Papers: Categorical Data Analysis and Experimental Design
Salon G	41. Contributed Papers: Designing Clinical Trials
Salon E	42. Contributed Papers: Competing Risks and Cure Rates
Salon C	43. Contributed Papers: Analyzing Microarray Data
Meeting Room 410	44. Contributed Papers: Statistical Methods in Genetics
6:00 p.m.–7:30 p.m.	President's Reception (By Invitation Only)
Salon H	

TUESDAY, MARCH 22

7:30 a.m.–5:00 p.m.	Conference Registration
Grand Ballroom Pre-function Area	
9:30 a.m.–4:00 p.m.	Placement Service
Meeting Rooms 406 & 408	
8:30 a.m.–5:30 p.m.	Exhibits Open
Grand Ballroom Pre-function Area	
8:30 a.m.–10:15 a.m.	Tutorial
Salon F	T4: Statistical Methods for Gel Electrophoresis Proteomics
	Scientific Program
Salon A	45. A Practicum on Multistate Survival Models
Salon B	46. Statistics in Disease Ecology
Meeting Rooms 402	47. Statistical Methods for Reproductive Epidemiology
Salon K	48. Modeling Brain Images—The Effects of Space, Time, and Individuality
Salon G	49. Crossover Designs for the Pharmaceutical Industry
Meeting Room 415	50. Contributed Papers: Diagnostic Tests
Meeting Room 410	51. Contributed Papers: Mixed Models: Linear, Generalized, and Non-Linear
Salon J	52. Contributed Papers: Bayesian Methods
Salon D	53. Contributed Papers: Missing Data Methods
Salon E	54. Contributed Papers: Methods for Multiple Endpoints
Meeting Room 412	55. Contributed Papers: Quantitative-Trait Linkage Analysis
10:15 a.m.–10:30 a.m.	Break
Grand Ballroom Pre-function Area	
10:30 a.m.–12:15 p.m.	Presidential Invited Address
Salon H	
12:30 p.m.–4:30 p.m.	Regional Committee (RECOM) Luncheon Meeting (By Invitation Only)
Meeting Room 404	

ENAR

Program Summary

1:45 p.m.–3:30 p.m.	Tutorial
Salon F	T5: Bayesian Approaches for Clinical Trial Design and Analysis
	Scientific Program
Salon A	56. Screening for Disease: Issues in Study Design and Analysis
Salon B	57. Combining Information Across Spatial Scales
Salon G	58. Bayesian Procedures for Analyzing Microarray Data
Salon D	59. Cost-Effectiveness Analysis: Methodologies for Comparing Competing Health Care Interventions (Theme Session)
Meeting Room 402	60. Recent Advances in Semiparametric Estimation
Meeting Room 415	61. Contributed Papers: Bioassay and Biopharmaceutical Applications
Salon E	62. Contributed Papers: Sequential Methods
Meeting Room 410	63. Contributed Papers: Survival Analysis II
Meeting Room 412	64. Contributed Papers: Longitudinal Data Analysis and Generalized Linear Models
Salon J	65. Contributed Papers: Bayesian Methods in Clinical Trials
Salon K	66. Contributed Papers: Bayesian Methods in Genomic Data Analysis
3:30 p.m.–3:45 p.m.	Break
Grand Ballroom Pre-function Area	
3:45 p.m.–5:30 p.m.	Scientific Program
Salon A	67. Methodologic Issues in Studies Involving Cancer Screening
Salon G	68. Current Advances in Modeling Time Course Gene Expression Data
Salon B	69. Statistical Issues in a Biometric Survey: The National Health and Nutrition Examination Survey (NHANES)
Salon D	70. Structural Equations and Psychometric Methods in Biological Studies (Theme Session)
Salon K	71. Seeking Pattern in Genomic Data Using Cross-Species Comparisons
Meeting Room 412	72. Contributed Papers: Semiparametric and Nonparametric Modeling
Meeting Room 415	73. Contributed Papers: Imaging of Brain Activity
Salon E	74. Contributed Papers: Cox Regression Models
Meeting Room 410	75. Contributed Papers: Longitudinal Data Analysis
Meeting Room 402	76. Contributed Papers: Topics in Biostatistics: From Quantal Responses to Case-Control Studies
Salon J	77. Contributed Papers: From Proteomics to Classification to Measurement Error: Current Topics in Proteomics/Genomics
5:30 p.m.–6:30 p.m.	ENAR Business Meeting (Open to All Members)
Meeting Room 415	
6:30 p.m.–9:30 p.m.	Tuesday Night Event (Heart of Texas Evening—Registration Required)
WEDNESDAY, MARCH 23	
7:30 a.m.–9:00 a.m.	Annual Meeting Planning Committee Meeting (Closed)
Board Room 401	
8:00 a.m.–12:15 p.m.	Conference Registration
Grand Ballroom Pre-function Area	
8:00 a.m.–12:00 noon	Exhibits Open
Grand Ballroom Pre-function Area	
8:30 a.m.–10:15 a.m.	Scientific Program
Salon A	78. Advanced Topics in Prostate Cancer Modeling: A Multidisciplinary and Integrated Perspective
Salon G	79. Statistical Methods in Quantitative Genetics and Genomics

Salon K	80. Statistical Methods for Analysis of Gene-Environment Interaction
Salon B	81. Isotonic Methods in Toxicology & Risk
Salon D	82. Novel Environmental Applications of Spatial Statistics
Meeting Room 415	83. Contributed Papers: Nonparametrics
Meeting Room 404	84. Contributed Papers: Statistical Methods in Screening
Salon E	85. Contributed Papers: Measurement Error
Salon J	86. Contributed Papers: Frailty Models
Salon F	87. Contributed Papers: Multiple Imputation
Meeting Room 412	88. Contributed Papers: Gene Expression Analysis in Cancer Research
10:15 a.m.–10:30 a.m.	Break
Grand Ballroom Pre-function Area	
10:30 a.m.–12:15 p.m.	Scientific Program
Salon A	89. Bayesian and Non-Bayesian Approaches to Competing Risks
Salon G	90. Integrating Multiple Sources of Genomic Data
Salon B	91. Biosurveillance GeoInformatics for BioSecurity (Theme Session)
Salon D	92. Risk Ranking and Disease Mapping
Salon E	93. To Mix or Not to Mix
Meeting Room 415	94. Contributed Papers: State Space Models and Time Series Analysis
Meeting Room 412	95. Contributed Papers: Causal Inference
Salon K	96. Contributed Papers: Clinical Trials II
Meeting Room 404	97. Contributed Papers: Spatial Modeling
Salon J	98. Contributed Papers: Survival Analysis III
Salon F	99. Contributed Papers: Associative Analysis of Genetic Data

ENAR

Scientific Program

Times may change slightly prior to the meetings. Please check the on-site program for final times. Asterisks (*) indicate paper presenters. Distinguished Student Paper Award presentations appear in **boldface.**

Monday, March 21
8:30–10:15 a.m.

1. RECENT DEVELOPMENTS IN SEQUENTIAL CLINICAL TRIALS METHODOLOGY *Salon G*

Sponsor: ASA Biopharmaceutical Section
Organizer: Linda J. Young, University of Florida
Chair: Linda J. Young, University of Florida

8:30 Dose Finding Based on Efficacy and Toxicity in Phase I/II Clinical Trials
Peter F. Thall* and John D. Cook, University of Texas M. D. Anderson Cancer Center

9:00 Design of Group Sequential Clinical Trials with Ordinal Categorical Data
Jonathan J. Shuster, Lili Tian*, and Myron Chang, University of Florida

9:30 Efficient Group Sequential Designs When There Are Several Effect Sizes Under Consideration
Chris Jennison, University of Bath, U.K.; Bruce W. Turnbull*, Cornell University

10:00 Floor Discussion

2. CUTTING-EDGE RISK ASSESSMENT ISSUES AND METHODS *Salon A*

Sponsor: ASA Risk Analysis Section
Organizer: Hojin Moon, NCTR/FDA
Chair: Ralph L. Kodell, NCTR/FDA

8:30 Developmental Neurotoxicity Modeling and Risk Assessment
Mehdi Razzaghi*, Bloomsburg University; Ralph L. Kodell, NCTR/FDA

8:55 Application of Risk Assessment and Modeling Approaches for Evaluating Gene Therapy Risks
Steven Anderson*, CBER/FDA

9:20 Comparing Model Averaging with Other Model Selection Strategies for Risk Estimation
Matthew W. Wheeler*, National Institute for Occupational Safety and Health; A. John Bailer, Miami University

9:45 Model Averaging Using the Kullback Information Criterion in Estimating Effective Doses for Microbial Infection and Illness
Hojin Moon*, Ralph L. Kodell, and James J. Chen, NCTR/FDA; Hyun-Joo Kim, Truman State University; David W. Gaylor, Gaylor and Associates, LLC

10:10 Floor Discussion

3. ANALYSIS AND ISSUES IN MATCHED FAMILY AGGREGATION STUDIES *Salon B*

Sponsor: ASA Section on Epidemiology
Organizer: John Williamson, Centers for Disease Control and Prevention
Chair: John Williamson, Centers for Disease Control and Prevention

8:30 Advances in the Design and Analysis of Familial Aggregation Studies
Abigail G. Matthews, Harvard University; Dianne M. Finkelstein, Massachusetts General Hospital; Rebecca A. Betensky*, Harvard University

9:00 Family-Specific Approaches to the Analysis of Case-Control Family Data
John Neuhaus*, University of California, San Francisco; Alastair Scott and Chris Wild, University of Auckland, New Zealand

9:30 A Semiparametric Method for Analyzing Matched Case-Control Family Studies
Molin Wang*, Harvard University and Dana-Farber Cancer Institute; John M. Williamson, National Center for Infectious Diseases, Centers for Disease Control and Prevention; Susan Redline, Rainbow Babies and Children's Hospital and Case Western Reserve University

10:00 Floor Discussion

4. *BIOMETRICS* SPECIAL INVITED PAPER SESSION *Salon C*

Sponsor: ENAR
Organizer: Xihong Lin, University of Michigan School of Public Health
Chair: Xihong Lin, University of Michigan School of Public Health

8:30 Recent Developments in Computational Biology
Wing Hung Wong* and Hongkai Ji, Stanford University

9:30 Discussants: Chip Lawrence, Brown University; Hongyu Zhao, Yale University

Scientific Program

5. LACK OF FIT TESTS FOR MODEL MISSPECIFICATION WITH APPLICATIONS TO LONGITUDINAL DATA AND SURVIVAL DATA ANALYSIS *Salon D*

Sponsor: IMS
Organizer: Annie Qu, Oregon State University
Chair: Runze Li, Pennsylvania State University

8:30 Using Bayesian Statistics to Test for Lack of Fit in Frequentist Fashion
Jeffrey D. Hart*, Texas A&M University

8:55 A Goodness-of-Fit-Test for Proportional Odds Models for Survival Data
Linxu Liu*, Columbia University; Jianhua Huang, University of Pennsylvania

9:20 The IOS Test for Model Misspecification
Brett Presnell* and Marinela Capanu, University of Florida; Dennis D. Boos, North Carolina State University

9:45 Goodness-of-Fit Test for Model Assumptions in Longitudinal Data
Annie Qu*, Oregon State University

10:10 Floor Discussion

6. CONTRIBUTED PAPERS:COMPUTATIONAL METHODS *Salon E*

Sponsor: ENAR
Chair: David Shera, University of Pennsylvania/CHOP

8:30 Partial Prior Improving NPML Estimation for Mixtures
Ji-Ping Z. Wang*, Northwestern University

8:45 Wavelets and Evolution Algorithms for Mass Spectrometry Data Processing
Ming Li*, Huiming Li, Johnathan Xu, and Yu Shyr, Vanderbilt University; Don Hong, East Tennessee State University and Vanderbilt University

9:00 Recent Progress on Mass Spectrometry Data Processing
Don Hong*, East Tennessee State University and Vanderbilt University; Yu Shyr, Vanderbilt University

9:15 Stochastic Simulation of E. coli 0157:H7 Infection in Cattle
Baktiar Hasan*, Brian Allen, and Scott A. McEwen, University of Guelph

9:30 Moment Estimators for Mutation Rates
Loki Natarajan*, Charles C. Berry, and Christoph Gasche, University of California, San Diego

9:45 A Unified Approach for Simultaneous Clustering and Differential Expression Identification
Ming Yuan*, Georgia Institute of Technology; Christina Kendziorski, University of Wisconsin–Madison

10:00 Modeling P-values in High-Dimensional Testing Applications Using a Uniform-Beta Mixture: The Performance of Interval Estimates
Qinfang Xiang and Gary L. Gadbury*, University of Missouri–Rolla; Jode Edwards, USDA ARS and Iowa State University

7. CONTRIBUTED PAPERS:NONPARAMETRIC METHODS IN LONGITUDINAL AND SURVIVAL ANALYSIS *Meeting Room 410*

Sponsor: ENAR
Chair: Sebastien Haneuse, Vanderbilt University

8:30 Nonparametric Tests for Dependent Observations Obtained at Varying Time Points
Susanne May*, University of California, San Diego; Victor DeGruttola, Harvard University

8:45 Nonparametric Regression Subject to a Monotonicity Constraint
Matthew J. Schipper*, Jeremy Taylor, and Xihong Lin, University of Michigan

9:00 On Minimax Wavelet Estimator with Censored Data
Linyuan Li*, University of New Hampshire

9:15 Inference for the Proportional Odds Model with a Change-Point Based on a Covariate Threshold
Rui Song* and Michael R. Kosorok, University of Wisconsin–Madison

9:30 On Kernel Function Estimation for Censored Data
Kagba N. Suaray*, University of California, Riverside

9:45 Penalized Likelihood Based Cross-validation Methods for Survival Data Analysis
Bin Wang*, University of South Alabama

10:00 Time-Varying Functional Regression for Predicting Remaining Lifetime Distributions from Longitudinal Trajectories
Hans-Georg Müller* and Ying Zhang, University of California, Davis

8. CONTRIBUTED PAPERS:SURVEY DATA METHODS
Meeting Room 415

Sponsor: ENAR
Chair: James Stamey, Stephen F. Austin State University

8:30 Can Population Estimates with Bridged-Race Categories Be Improved Using the Census Quality Survey?
Deborah D. Ingram*, National Center for Health Statistics

8:45 An Approach to Estimating the Distribution and Bias of Comparative Fit Indices Using Binary Data
Zara E. Sadler*, Barbara C. Tilley, Philip F. Rust, and Peng Huang, Medical University of South Carolina; Linda M. Kaste, University of Illinois at Chicago

9:00 Finite Population Cumulative Distribution Functions and Measurement Error
Jeremy Aldworth*, RTI International

9:15 Model-Based Estimates of the Finite Population Mean for Two-Stage Cluster Samples with Unit Nonresponse
Ying Yuan* and Roderick J. A. Little, University of Michigan

9:30 Unbalanced Ranked Set Sampling for Estimating a Population Proportion Under Imperfect Rankings
Haiying Chen*, Wake Forest University School of Medicine; Elizabeth Stasny and Douglas Wolfe, Ohio State University

9:45 Estimation of Prevalence of Overweight in Small Areas— A Robust Extension of the Fay-Herriot Model
Dawei Xie*, University of Pennsylvania; Trivellore E. Raghunathan and James M. Lepkowski, University of Michigan

10:00 Jackknife Variance Estimation of the Regression and Calibration Estimator for Two 2-Phase Samples
Jong-Min Kim* and Jon E. Anderson, University of Minnesota, Morris

9. CONTRIBUTED PAPERS: ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS *Meeting Room 412*

Sponsor: ENAR
Chair: Ronald Gangnon, University of Wisconsin–Madison

8:30 Hierarchical Bayesian Galerkin-Based Parameterizations of Spatio-Temporal Dynamical Models with Application to Ecological Processes
Ali Arab* and Christopher K. Wikle, University of Missouri–Columbia

8:45 Diagnostic Approaches to Statistical Models Incorporating Dynamic Biological Components
Michael B. Brimacombe*, New Jersey Medical School– UMDNJ

9:00 A Hierarchical Bayesian Approach for Describing the Spatio-Temporal Dynamics of Invasive Species
Mevin B. Hooten* and Christopher K. Wikle, University of Missouri–Columbia

9:15 Test for Independence Between Marks and Points of a Marked Point Process
Yongtao Guan*, University of Miami

9:30 Combining Information from Multiple Sources to Estimate the Probability of a Rare Event
Philip M. Dixon*, Iowa State University

9:45 Tests for Order-Restrictions in Ordinal Data: A Graphical Approach
Eric R. Teoh*, University of North Carolina at Chapel Hill; Abraham Nyska, National Institute of Environmental Health Sciences; Uri Wormser, Hebrew University of Jerusalem; Shyamal D. Peddada, National Institute of Environmental Health Sciences

10:00 Some Challenges Encountered When Analyzing Biomarker Data
Stephen W. Looney* and Joseph L. Hagan, Louisiana State University Health Sciences Center School of Public Health

10. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS I
Salon J

Sponsor: ENAR
Chair: Daniel J. Sargent, Mayo Clinic

8:30 Estimating Survival Distributions in the Presence of Informative Censoring via a Latent Survival Time Imputation Approach
Pai-Lien Chen*, Family Health International; Bosny Pierre-Louis, Family Health International and University of North Carolina at Chapel Hill

8:45 Variable Selection for Censored Data
Brent A. Johnson*, University of North Carolina at Chapel Hill

9:00 A Simple Approach to the Estimation of the Survival Function Based on Two-Stage Sampling for the Dependent Censorship
Seungyeoun Lee*, Sejong University–Seoul, Korea

9:15 **Nonparametric Estimation of the Concordance Correlation Coefficient Under Univariate Censoring**
Ying Guo* and Amita K. Manatunga, Rollins School of Public Health, Emory University

9:30 Constructing Exact Confidence Bounds for the True Survival Curve Using the Kaplan-Meier Survival Function
Craig B. Borkowf*, Centers for Disease Control and Prevention

9:45 Testing Goodness-of-Fit of a Truncation Model
Micha Mandel* and Rebecca Bentensky, Harvard School of Public Health

10:00 A Pearson Goodness-of-Fit Test for Interval Censored Data
Denise Babineau* and Jerry Lawless, University of Waterloo

11. ANALYZING HIGH-DIMENSIONAL GENOMIC DATA *Salon K*

Sponsor: ENAR
Chair: Dirk Moore, University of Medicine and Dentistry, New Jersey

8:30 Prediction Error Estimation: A Comparison of Resampling Methods
Annette M. Molinaro*, Ruth Pfeiffer, and Richard Simon, National Cancer Institute

8:45 Bayesian Identification of Prognostic Molecular Signatures for Survival Phenotypes
Dabao Zhang*, University of Rochester Medical Center

9:00 Using Longitudinal Genomic Data to Predict Failure Outcomes
Natasa Rajicic*, Dianne Finkelstein, and David Schoenfeld, Harvard University

9:15 Statistical Models for Characterizing Fundamental Patterns Underlying Gene Expression Profiles
Fei Long*, Tian Liu, and Rongling Wu, University of Florida

9:30 Feature-Specific Constrained Latent Class Analysis for Genomic Data
Andres Houseman*, Brent A. Coull, and Rebecca A. Betensky, Harvard University

9:45 Sharp Simultaneous Intervals for the Means of Selected Populations with Application to Microarray Data Analysis
Jing Qiu*, University of Missouri–Columbia; Gene J. T. Hwang, Cornell University

10:00 Floor Discussion

Monday, March 21
10:15–10:30 a.m.

Break *Grand Ballroom Pre-Function Area*

Monday, March 21
10:30 a.m.–12:15 p.m.

12. RECENT ADVANCES IN THE ANALYSIS OF RECURRENT EVENTS DATA *Salon A*

Sponsors: ENAR/ASA Risk Analysis Section
Organizer: Xuelin Huang, University of Texas M. D. Anderson Cancer Center
Chair: Xuelin Huang, University of Texas M. D. Anderson Cancer Center

10:30 Recurrent Events and Longitudinal Markers
Edsel A. Pena*, Elizabeth H. Slate, and Jun Han, University of South Carolina

10:55 Semiparametric Inference of Successive Durations
Yijian Huang*, Emory University

11:20 Frailty in the Accelerated Gap Times Model
Robert L. Strawderman*, Cornell University

11:45 Analyzing Recurrent Longitudinal Data with Complication of Informative Censoring
Mei-Cheng Wang*, Johns Hopkins University

12:10 Floor Discussion

13. NATURAL RESOURCE ESTIMATION FOR SMALL AREAS (THEME SESSION) *Salon B*

Sponsor: ASA Section on Statistics and the Environment
Organizer: Ronald E. McRoberts, USDA Forest Service
Chair: Mary C. Christman, University of Florida

10:30 Nonparametric Small Area Estimation Using Penalized Spline Regression
Jean Opsomer*, Iowa State University; Jay Breid, Colorado State University; Gerda Claeskens, Katholieke Universiteit Leuven (Belgium); Goeran Kauermann, Universitaet Bielefeld (Germany); Giovanna Ranalli, Universita di Perugia (Italy)

10:55 Small Area Estimation of Forest Attributes for a Temporally Continuous Sampling Design
Francis A. Roesch*, USDA Forest Service

11:20 Map-Based Estimation of Forest Area
Ronald E. McRoberts*, USDA Forest Service

11:45 Estimating Coho Salmon Abundance in Small Stream Basins Using a Space-Time Model with Covariates
Don L. Stevens Jr.* and Ruben A. Smith, Oregon State University

12:10 Floor Discussion

Scientific Program

14. DEALING WITH MALDI-TOF/SELDI-TOF PROTEOMIC DATA: EXPERIMENTAL DESIGN AND LOW-LEVEL PROCESSING *Salon J*

Sponsor: ENAR
Organizer: Keith A. Baggerly, University of Texas M. D. Anderson Cancer Center
Chair: Jeffrey S. Morris, University of Texas M. D. Anderson Cancer Center

10:30 MALDI-TOF/SELDI-TOF: Background and Design Issues
Keith A. Baggerly* and Jeffrey S. Morris, University of Texas M. D. Anderson Cancer Center

11:00 Enhancement of Sensitivity and Resolution of Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometric Records
Dariya I. Malyarenko*, William E. Cooke, Eugene R. Tracy, Michael W. Trosset, Haijian Chen, Dennis M. Manos, College of William and Mary; Maciek Sasinowski, INCOGEN, Inc.; John Semmes and Gunjan Malik, Eastern Virginia Medical School

11:30 Denoising of Mass Spectrometry Data: Wavelets and Averaging
Jianhua Hu*, Kevin R. Coombes, Keith A. Baggerly, and Jeffrey S. Morris, University of Texas M. D. Anderson Cancer Center

12:00 Floor Discussion

15. AT THE CROSSROADS OF BIOSTATISTICS AND SECURITY BIOMETRICS (THEME SESSION) *Salon D*

Sponsors: ENAR/ASA Section on Statistics in Defense and National Security
Organizer: Peter B. Imrey, Cleveland Clinic Foundation
Chair: Peter B. Imrey, Cleveland Clinic Foundation

10:30 (Security) Biometrics from a (Statistical) Biometrics Perspective
Michael E. Schuckers*, St. Lawrence University and Center for Identification Technology Research

11:00 Statistical Issues in Biometric Identification
David Banks*, Duke University

11:30 The DNA Testing Experience
Bruce S. Weir*, North Carolina State University

12:00 Discussant: Peter B. Imrey, Cleveland Clinic Foundation

16. NEW DEVELOPMENTS IN THE ANALYSIS OF HIGH-DIMENSIONAL DATA *Salon C*

Sponsor: ASA Biometrics Section
Organizer: Glen Satten, Centers for Disease Control and Prevention
Chair: Glen Satten, Centers for Disease Control and Prevention

10:30 Prediction by Supervised Principal Components
Brad Efron, Trevor Hastie, Ian Johnstone, and Rob Tibshirani*, Stanford University

11:10 The Entire Regularization Path for the Support Vector Machine
Trevor Hastie*, Saharon Rosset, Rob Tibshirani, and Ji Zhu, Stanford University

11:50 Discussant: Bob Stine, University of Pennsylvania

17. CONTRIBUTED PAPERS:HEALTH SERVICES RESEARCH *Meeting Room 415*

Sponsor: ENAR
Chair: Alka Indurkhya, Harvard University

10:30 Bootstrap Confidence Bands
Laura L. Johnson*, National Institutes of Health; Paula Diehr, University of Washington

10:45 Simulation of Breast Cancer Screening Among Women Veterans
Wenyaw Chan*, David R. Lairson, David P. Smith, and Yen-Peng Li, University of Texas Health Science Center at Houston School of Public Health

11:00 Estimating Incremental Cost-Effectiveness Ratios and Their Confidence Intervals with Differentially Censored Data
Hongkun Wang* and Hongwei Zhao, University of Rochester

11:15 Modeling Infant Birthweight
Katherine J. Hoggatt*, Sander Greenland, and Beate R. Ritz, University of California, Los Angeles

11:30 Estimating the Effectiveness of Mental Health Services in a Study of the 1998 U.S. Embassy Bombing in Nairobi, Kenya
Haekyung Jeon-Slaughter* and Betty Pfefferbaum, University of Oklahoma Health Sciences Center; Carol S. North, Washington University School of Medicine; Pushpa Narayanan, University of Oklahoma Health Sciences Center; Lee Ann Ross, United States Agency for International Development, Kenya

11:45 Conversion of Two Continuous Measures Using Truncated Regression: Barthel and FIM Scores of Stroke Patients
Yongsung Joo*, University of Florida and V.A. Hospital; Keunbaik Lee and George Casella, University of Florida; Sooyeon Kim and Pamela Duncan, V.A. Hospital

12:00 **Analyzing the Effect of Covariates on Gap Times Between Interval-Censored Recurrent Events**
Dan Sheng, New York University and Mimi Kim, Albert Einstein College of Medicine of Yeshiva University

18. CONTRIBUTED PAPERS: MODELING METHODS IN EPIDEMIOLOGY *Salon E*

Sponsor: ENAR
Chair: Charles Hall, Albert Einstein College of Medicine of Yeshiva University

10:30 Modeling Prostate Cancer Incidence: Racial Differences
Aniko Szabo* Huntsman Cancer Institute; Alexander D. Tsodikov, University of California, Davis

10:45 A Bayesian Hierarchical Model for Estimation of Disease Incidence Using Two Surveillance Data Sets
Joan Buenconsejo*, Yale University/Food and Drug Administration; Durland Fish and Theodore Holford, Yale University; James E. Childs, Yale University/ Centers for Disease Control and Prevention

11:00 Genetic Misclassification in Case-Control Studies
Christie M. Spinka*, University of Missouri–Columbia; Nilanjan Chatterjee, National Cancer Institute; Raymond J. Carroll, Texas A&M University

11:15 Modeling Hormone Reproductive Data Using Complementary Approaches
Irina Bondarenko* and MaryFran Sowers, University of Michigan; Daowen Zhang, North Carolina State University

11:30 The Impact of the Analytic Approach on Coronary Calcium and Cardiovascular Risk Factors
Imke Janssen*, Zhen Chen, and Lynda H. Powell, Rush University Medical Center

11:45 Developing Indices of Disease Severity in Mortality Prediction
Guofen Yan*, Tom Greene, and Gerald Beck, Cleveland Clinic Foundation

12:00 Statistical Modeling and Inference for HIV/AIDS
Sujay Datta*, Northern Michigan University

19. CONTRIBUTED PAPERS: ENVIRONMENTAL AND TOXICOLOGICAL APPLICATIONS *Meeting Room 410*

Sponsor: ENAR
Chair: Yongtao Guan, University of Miami

10:30 Detecting Departure from Additivity Along a Fixed-Ratio Ray with a Piecewise Model for Dose and Interaction Thresholds
Sharon D. Yeatts* and Chris Gennings, Virginia Commonwealth University; Timothy E. O'Brien, Loyola University, Chicago

10:45 Using a Bayesian Hierarchical Model to Estimate the Rate of Emission of Greenhouse Gases from a Facility Housing Pigs
Cory R. Heilmann* and Philip Dixon, Iowa State University

11:00 Spatial Estimation of Risk of Mortality Due to Air Pollution
Hae-Ryoung Song*, Montserrat Fuentes, and Sujit Ghosh, North Carolina State University

11:15 Threshold Regression and Applications in Environmental Research
Mei-Ling T. Lee*, Harvard University; George A.Whitmore, McGill University

11:30 Estimating Chronic Effects of Fine Particles (PM2.5) on Adult Mortality at Different Spatial and Temporal Scales
Sorina E. Eftim*, Francesca Dominici, Aidan McDermott, Scott L. Zeger, and Jonathan M. Samet, Johns Hopkins University

11:45 Data Analysis and Methods for the Study of Environmental Impacts of Rocket Emissions
Yi Ye* and Gary L. Gadbury, University of Missouri–Rolla

12:00 Empirical Evaluation of Sufficient Similarity in Inference for a Mixture of Many Chemicals Using a Fixed-Ratio Ray Design
LeAnna G. Stork*, Chris Gennings, W. Hans Carter Jr., Robert E. Johnson, and Darcy P. Mays, Virginia Commonwealth University

20. CONTRIBUTED PAPERS: ADAPTIVE METHODS AND DESIGNS *Salon G*

Sponsor: ENAR
Chair: Andy Willan, Hospital for Sick Children

10:30 Adaptive Bayes Designs for Dose-Finding Phase I Clinical Trials
Yisheng Li, Yuan Ji*, and Benjamin Bekele, University of Texas M. D. Anderson Cancer Center

10:45 **Adaptive Design for Censored Survival Data Adjusting for Covariates**
Jie Yang*, University of Florida; Pei-Yun Chen and Kaifeng Lu, Merck Research Laboratories

11:00 Adaptive Model-Based Designs for Dose-Finding Studies
Vladimir Dragalin* and Valerii Fedorov, GlaxoSmithKline

11:15 Adaptive Two-Stage Designs in Phase II Clinical Trials
Anindita Banerjee* and Anastasios A. Tsiatis, North Carolina State University

11:30 An Adaptive Single-Step FDR Procedure with Applications to DNA Microarray Analysis
Vishwanath Iyer*, Bristol Myers Squibb; Sanat Sarkar, Temple University

11:45 When Are Adaptive Designs Appropriate?
Cyrus R. Mehta*, Cytel Software Corporation

12:00 Floor Discussion

21. CONTRIBUTED PAPERS: ANALYSIS OF CORRELATED DATA *Meeting Room 412*

Sponsor: ENAR
Chair: Tom Tenhave, University of Pennsylvania

10:30 The Simultaneous Analysis of Mixed Types of Outcomes Using Nonlinear Threshold Models
Todd Coffey* and Chris Gennings, Virginia Commonwealth University

10:45 Goodness-of-Fit Tests for Binomial Generalized Estimating Equations (GEE) Models: Simulation Results
Huiyi Lin* and Leann Myers, Tulane University

11:00 Estimating Equation Approach for Truncated Covariates
Gina M. D'Angelo* and Lisa Weissfeld, University of Pittsburgh Graduate School of Public Health

11:15 Comparison of Wang-Carey Estimation Versus Quasi-Least Squares
Wenguang Sun* and Justine Shults, University of Pennsylvania School of Medicine

11:30 Adjusted Quasi-Least Squares for Valid Analysis of Correlated Binary Data
Justine Shults* and Wenguang Sun, University of Pennsylvania School of Medicine

11:45 Estimation of Clustered Poisson Regression with Random Intercepts
Eugene Demidenko*, Dartmouth Medical School

12:00 Floor Discussion

22. CONTRIBUTED PAPERS: LINKAGE ANALYSIS *Salon K*

Sponsor: ENAR
Chair: Ken Hess, University of Texas M. D. Anderson Cancer Center

10:30 On Family-Based Genetic Analysis Allowing for Missing Parental Information
Jing Han* and Yongzhao Shao, New York University School of Medicine

10:45 Linkage Analysis of Ordinal Traits for Pedigree Data
Rui Feng*, James F. Leckman, and Heping Zhang,Yale University

11:00 **Joint Modeling of Linkage and Association: Identifying SNPs Responsible for a Linkage Signal**
Mingyao Li*, Michael Boehnke, and Goncalo R. Abecasis, University of Michigan

11:15 Detection of Pleiotropic Genetic Effects in Quantitative Multivariate Linkage Analysis
Mariza de Andrade*, Curtis Olswold, and Stephen T. Turner, Mayo Clinic College of Medicine

11:30 Model Free Linkage Analysis When the Number of Alleles Shared Identical by Descent Between Relative Pairs Is Missing
Tao Wang* and Robert C. Elston, Case Western Reserve University

11:45 Combining Evidence from Linkage and Association Studies Using Dempster-Shafer Theory
Chun Li* and Dan Hahs, Vanderbilt University

12:00 A Logistic Regression Mixture Model for Interval Mapping of Genetic Trait Loci Affecting Binary Phenotypes
Weiping Deng*, George Washington University; Hanfeng Chen, Bowling Green State University; Zhaohai Li, George Washington University

Monday, March 21
12:15–1:30 p.m.

ROUNDTABLE LUNCHEONS *Salon H*
(see pages 16-20)

Monday, March 21
1:45–3:30 p.m.

23. STATISTICAL ISSUES AND NOVEL METHODS IN VACCINE CLINICAL TRIALS *Salon A*

Sponsor: ASA Section on Epidemiology/ASA Biopharmaceutical Section
Organizer: William W. B. Wang, Merck Research Laboratories
Chair: William W. B. Wang, Merck Research Laboratories

1:45 Evaluating HIV Vaccine: Selecting Endpoints and Validating Surrogates
Thomas Fleming*, University of Washington

2:15 Causal Vaccine Effects on Binary Post-Infection Outcomes
Michael G.Hudgens, University of North Carolina at Chapel Hill; M. Elizabeth Halloran*, Emory University

2:45 Demonstrating That an HIV Vaccine Lowers the Risk and/or Severity of HIV Infection
Devan V. Mehrotra* and Xiaoming Li, Merck Research Laboratories; Peter B. Gilbert, Fred Hutchinson Cancer Research Center

3:15 Discussant: Constantine Frangakis, Johns Hopkins University

24. STATISTICAL ANALYSIS OF WILDFIRE DATA (THEME SESSION) *Salon K*

Sponsors: ASA Section on Statistics and the Environment/ASA Risk Analysis Section
Organizer: Marcia Gumpertz, North Carolina State University
Chair: Marcia Gumpertz, North Carolina State University

1:45 Estimating Wildfire Management Effectiveness
David T. Butry*, USDA Forest Service; Marcia L. Gumpertz, North Carolina State University; Marc G. Genton, Texas A & M University

2:10 Towards Improved Prediction of Wildfire Risk
Frederic P. Schoenberg*, Maria Chang, Haiyong Xu, and Jamie Pompa, University of California, Los Angeles; James Woods, California State University, Long Beach; Roger D. Peng, Johns Hopkins University

2:35 Fire Regimes: Controls at Different Scales of Space and Time
Max A. Moritz*, Ecosystem Sciences, ESPM, University of California, Berkeley

3:00 Modeling Size of Large Catastrophic Wildfires Using Skew-Elliptical Distributions
Marc G. Genton*, Texas A&M University

3:25 Floor Discussion

25. BAYESIAN STATISTICAL MODELING OF MASS SPECTROMETRY PROTEOMIC DATA *Salon C*

Sponsor: ENAR
Organizer: Jeffrey S. Morris, University of Texas M. D. Anderson Cancer Center
Chair: Keith Baggerly, University of Texas M. D. Anderson Cancer Center

1:45 Sources of Variability in MALDI-TOF MS Protein Profiling
Dean Billheimer*, Vanderbilt University

2:10 A Bayesian Mixture Model for Protein Biomarker Discovery
Kim-Anh Do* and Peter Mueller, University of Texas M. D. Anderson Cancer Center; Raj Bandyopadhya, Rice University

2:35 Bayesian Nonparametric Models for Proteomic Expression
Merlise A. Clyde*, Leanna House, and Robert Wolpert, Duke University

3:00 Bayesian Modeling and Inference for Mass Spectrometry Data Using Functional Mixed Models
Jeffrey S. Morris*, Kevin R. Coombes, and Keith A. Baggerly, University of Texas M. D. Anderson Cancer Center; Philip J. Brown, University of Kent, Canterbury

3:25 Floor Discussion

26. NEW APPROACHES TO STATISTICAL ACCESS TO DATA IN A CONFIDENTIAL WORLD (THEME SESSION) *Salon B*

Sponsors: ASA Section on Statistics in Defense and National Security/ASA Section on Survey Research Methods/IMS
Organizer: Stephen E. Fienberg, Carnegie Mellon University
Chair: Stephen E. Fienberg, Carnegie Mellon University

1:45 Regression on Distributed Databases via Secure Multi-Party Computation
Alan F. Karr*, Xiaodong Lin, and Ashish P. Sanil, National Institute of Statistical Sciences; Jerome P. Reiter, Duke University

2:15 Bounds for Cell Entries in Multi-way Tables Given Combinations of Marginals and Conditionals
Aleksandra B. Slavkovic*, Pennsylvania State University

2:45 Algebraic Geometry Tools for Statistical Disclosure Limitation and Statistical Estimation in Contingency Tables
Alessandro Rinaldo and Stephen E. Fienberg*, Carnegie Mellon University; Aleksandra B. Slavkovic, Pennsylvania State University; Seth Sullivant, University of California, Berkeley

3:15 Floor Discussion

27. NON-NORMAL RANDOM EFFECTS MODELS *Salon D*

Sponsor: ENAR
Organizer: Peter Song, University of Waterloo
Chair: Grace Y. Yi, University of Waterloo

1:45 Non-normal Random Effects in Generalized Linear Mixed Models
Alan Agresti*, University of Florida

2:10 Partially Observed Information and Inference About Non-Gaussian Mixed Linear Models
Jiming Jiang*, University of California, Davis

2:35 A Semiparametric Likelihood Approach to Generalized Linear Models with Covariates as Random Effects for Longitudinal Data
Erning Li*, Texas A&M University; Daowen Zhang and Marie Davidian, North Carolina State University

3:00 Maximum Likelihood Inference in Non-Normal Random Effects Models
Peter X. K. Song* and Peng Zhang, University of Waterloo; Annie P. Qu, Oregon State University

3:25 Floor Discussion

28. CONTRIBUTED PAPERS: METHODS IN EPIDEMIOLOGY *Salon E*

Sponsor: ENAR
Chair: Sujay Datta, Northern Michigan

1:45 Robust Trend Tests for Genetic Association Using Matched Case-Control Design
Gang Zheng and Xin Tian*, National Heart, Lung, and Blood Institute

2:00 Education Delays Accelerated Decline in Memory in Persons Who Develop Alzheimer's Disease
Charles B. Hall*, Carol A. Derby, Mindy J. Katz, Aaron J. LeValley, Joe Verghese, Herman Buschke, and Richard B. Lipton, Albert Einstein College of Medicine of Yeshiva University

2:15 Population Lab: The Creation of Virtual Populations in Genetic Epidemiology Research
Monica Nichifor* and Marie Reilly, Karolinska Institutet, Stockholm, Sweden

2:30 Attributable Risk Estimation in Longitudinal Studies with Censoring
Cynthia S. Crowson*, Terry M. Therneau, Sherine E. Gabriel, and William M. O'Fallon, Mayo Clinic

2:45 Hierarchical Models for Combining Ecological and Case-Control Data
Sebastien J. Haneuse*, Vanderbilt University; Jonathan C. Wakefield, University of Washington

3:00 Causal Inference for Morbidity Outcomes in the Presence of Death
Brian L. Egleston*, Daniel O. Scharfstein, Ellen E. Freeman, and Sheila K. West, Johns Hopkins University

3:15 Floor Discussion

29. CONTRIBUTED PAPERS: CLINICAL TRIALS I *Salon G*

Sponsor: ENAR
Chair: Yuan Ji, University of Texas M. D. Anderson Cancer Center

1:45 A Sequential Procedure for Monitoring Clinical Trials Against Historical Controls
Xiaoping Xiong* , St. Jude Children's Research Hospital; Ming Tan, University of Maryland Greenebaum Cancer Center ; James Boyett, St. Jude Children's Research Hospital

2:00 Statistical Approaches for Evaluating Non-Inferiority Trials When the Non-Inferiority Margin Depends on the Control Event Rate
Mimi Y. Kim* and Xiaonan Xue, Albert Einstein College of Medicine of Yeshiva University

2:15 Predicting Event Times in Clinical Trials in the Absence of Treatment Arm Information
Mark Donovan*, Michael R. Elliott, and Daniel F. Heitjan, University of Pennsylvania

2:30 Simple Confidence Bounds at Effective Doses in Dose Finding Studies
Yi-Hsuan Tu* and Ying-Kuen Cheung, Columbia University

2:45 Equivalence Assessment on Multiple Proportion Outcomes
Lan Kong*, University of Pittsburgh; Robert C. Kohberger and Gary G. Koch, University of North Carolina at Chapel Hill

3:00 Incorporating Interim Analyses and/or Historical Controls into Flexible Screening Trials
Daniel J. Sargent*, Susan Geyer, and Haolan Lu, Mayo Clinic, Rochester

3:15 Quantifying Placebo Effect in Discontinuation Trials Using Functional Data
Eva Petkova* , Columbia University; Thaddeus Tarpey, Wright State University; Todd R. Ogden, Columbia University

30. CONTRIBUTED PAPERS: SPATIAL MODELING OF DISEASE *Meeting Room 415*

Sponsor: ENAR
Chair: Imke Janssen, Rush University Medical Center

1:45 Determinants of Small Area Racial Disparities in Stroke Mortality in the Southeastern United States, 1999–2003
Eric C. Tassone* , Emory University; Michele Casper, Centers for Disease Control and Prevention ; Lance A. Waller, Emory University; Ish Williams, Centers for Disease Control and Prevention; Katrina Moore, University of Washington and Centers for Disease Control and Prevention

2:00 Impact of Prior Choice on Localized Bayes Factors for Cluster Detection
Ronald E. Gangnon*, University of Wisconsin–Madison

2:15 Spatio-Temporal Analysis of Emergency-Room Visits for Ischemic Heart Disease in NSW, Australia
Sandy Burden, University of Wollongong, Australia; Subharup Guha*, Harvard University; Geoff Morgan, University of Sydney, Australia; Louise Ryan, Harvard University; Ross Sparks, Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia; Linda Young, University of Florida

2:30 **Detecting Spatial Clustering in Matched Case-Control Studies**
Andrea J. Cook* and Yi Li , Harvard University

2:45 **Spatial Analysis of Periodontal Data Using Conditionally Autoregressive Priors Having Two Types of Neighbor Relations**
Brian J. Reich*, James S. Hodges, and Bradley P. Carlin, University of Minnesota

3:00 Bayesian Cluster Modeling for Space-Time Disease Counts
Ping Yan* and Murray K. Clayton, University of Wisconsin–Madison
3:15 A State Space Model for Stochastic Gompertz Growth of Cancer Tumors
Wai-Yuan Tan and Weiming Ke*, University of Memphis

31. CONTRIBUTED PAPERS: RECURRENT EVENTS ANALYSIS *Meeting Room 410*

Sponsor: ENAR
Chair: Brent Johnson, University of North Carolina at Chapel Hill

1:45 Robust Estimation in Semiparametric Transformation Models for Censored Point Processes
Rajeshwari Sundaram*, University of North Carolina at Chapel Hill
2:00 Use of the Andersen-Gill Model to Evaluate Treatment Effect in the Presence of Disease Progression
Alexander C. Cambon*, University of Louisville
2:15 Semi-Parametric Regression for Recurrent Event Data with Time-Dependent Covariates and Informative Censoring
Xianghua Luo* and Mei-Cheng Wang, Johns Hopkins University
2:30 A General Class of Parametric Models for Recurrent Event Data
Russell S. Stocker*, Mississippi State University; Edsel A. Pena, University of South Carolina
2:45 **Estimating the Quality-of-Life-Adjusted Gap Time Distribution of Successive Events Subject to Censoring**
Adin-Cristian Andrei* and Susan Murray, University of Michigan
3:00 Analysis of Recurrent Events in the Pediatric Firearm Victim's Emergency Department Visit Database
Hyun J. Lim*, Marlene Melzer-Lange, and Liu Jingxia, Medical College of Wisconsin
3:15 Statistical Analysis of Panel Count Data
Do-Hwan Park and Jianguo Sun, University of Missouri; Xingqui Zhao, University of Alberta

32. CONTRIBUTED PAPERS: MISSING DATA IN LONGITUDINAL DATA ANALYSIS *Meeting Room 412*

Sponsor: ENAR
Chair: Jong-Min Kim, University of Minnesota, Morris

1:45 Latent Class Regression Models for Incomplete Longitudinal Binary Responses
I. Dunsiger* and Joseph W. Hogan, Brown University
2:00 A Latent Class Model for Longitudinal Binary Response Data with Nonignorable Missingness
Li Qin* , University of Pittsburgh; Changyu Shen, Indiana University; Lisa A. Weissfeld, University of Pittsburgh
2:15 Robust Methods for Longitudinal Data with Missing Observations
Grace Yi*, University of Waterloo; Wenqing He, University of Western Ontario
2:30 A Model for Incomplete Longitudinal Multivariate Ordinal Data
Li C. Liu*, Institute of Health Research and Policy, University of Illinois at Chicago
2:45 Sensitivity Analysis and Informative Priors for Longitudinal Binary Data with Outcome-Related Dropout
Joo Yeon Lee* and Joseph W. Hogan, Brown University
3:00 Classification of Multivariate Repeated Measures Data with Missing Values
Anuradha Roy*, University of Texas at San Antonio
3:15 Some Practical Solutions to Analyzing Messy Data
Monnie McGee*, Southern Methodist University

33. CONTRIBUTED PAPERS: MULTIPLE TESTING AND FALSE DISCOVERY RATES *Salon J*

Sponsor: ENAR
Chair: Joshua M. Tebbs, Kansas State University.

1:45 Profile Significance: A Powerful Alternative to False Discovery Rate Control
Cheng Cheng*, St. Jude Children's Research Hospital
2:00 Fast Estimation of Sample Size While Controlling for FDR in Multiple Testing
J. T. G. Hwang and Peng Liu*, Cornell University
2:15 Parametric and Nonparametric FDR Estimation Revisited
Baolin Wu*, University of Minnesota; Zhong Guan, Indiana University, South Bend; Hongyu Zhao, Yale University
2:30 The Faulty False Discovery Rate: Addressing Bias in Null P-Values to Adjust FDR Calculation
Hoa Nguyen*, Kathryn Roeder, and Larry Wasserman, Carnegie Mellon University
2:45 P-Values-Only-Based Stepwise Procedures for Multiple Testing and Their Optimality Properties
Alexander Y. Gordon*, University of Rochester

3:00 Finite-Sample Control of the Family-Wise Error Rate in Multiple Hypothesis Testing
Greg DiRienzo*, Harvard University
3:15 A Bayesian Analysis of Multiple Hypothesis Tests
Anthony L. Almudevar*, University of Rochester

Monday, March 21
3:30–3:45 p.m.

Break *Grand Ballroom Pre-Function Area*

Monday, March 21
3:45–5:30 p.m.

34. COMBINING SPATIAL MEASUREMENT PARAMETERS AND DESIGN OF EXPERIMENTS TO EVALUATE THE EFFECTIVENESS OF TREATMENTS FOR PRECISION AGRICULTURE (THEME SESSION) *Salon A*

Sponsor: ASA Section on Statistics and the Environment
Organizer: George Milliken, Kansas State University
Chair: George Milliken, Kansas State University

3:45 Using Spatial Information in Precision Agriculture Management of Cotton
Jeff Willers* and Chuck O'Hara, Mississippi State University; George Milliken, Kansas State University
4:15 Assembling the Spatial Information to Provide a Data Set for Statistical Analyses That Can be Used to Evaluate Precision
Chuck O'Hara* and Jeff Willers, Mississippi State University; George Milliken, Kansas State University
4:45 Using Spatial Information in the Design and Analysis of Experiments Used to Evaluate the Effectiveness of Precision Agriculture Management Practices
George Milliken*, Kansas State University; Jeff Willers and Chuck O'Hara, Mississippi State University
5:15 Floor Discussion

35. SPATIAL EPIDEMIOLOGY *Salon B*

Sponsor: ASA Section on Epidemiology
Organizer: Linda J. Young, University of Florida
Chair: Linda J. Young, University of Florida

3:45 Voronoi Tesselation in the Analysis of Small Area Cancer Incidence and Uranium in Ground Water
Fedele Greco*, University of Bologna, Italy; Andrew B. Lawson, University of South Carolina
4:15 Effects of Model Misspecification in the Analysis of Spatial Data
Louise M. Ryan*, Harvard Unversity
4:45 The Limitations of Spatial Analysis
Geoffrey M. Jacquez*, BioMedware and TerraSeer
5:15 Floor Discussion

36. PANEL: STATISTICAL CONCERNS UNDER THE FEDERAL ADVISORY COMMITTEE ACT (FACA) *Salon J*

Sponsor: ENAR
Organizer: Mary A. Foulkes, OBE/CBER/FDA
Chair: Mary A. Foulkes, OBE/CBER/FDA

3:45 Introduction
Mary A. Foulkes*, OBE/CBER/FDA
3:50 Panel Discussion
Greg Campbell, Center for Devices and Radiological Health, FDA; Christopher J. Portier, National Institute of Environmental Health Sciences; Janet Wittes, Statistics Collaborative; Fritz J. Scheuren, University of Chicago

37. STATISTICAL ANALYSIS OF DIFFUSION TENSOR IMAGING *Meeting Room 415*

Sponsor: ENAR
Organizers: Hongtu Zhu, Columbia University and Heping Zhang, Yale University
Chair: Hongtu Zhu, Columbia University

3:45 Statistical Analysis of Noise Models in Diffusion Tensor Images
Hongtu Zhu* and Dongrong Xu, Columbia College of Physicians and Surgeons; Heping Zhang, Yale University; Bradley Peterson, Columbia College of Physicians and Surgeons
4:10 Regularization of Diffusion Tensor Brain Images Via the Kernel Method
Moo Chung, University of Wisconsin-Madison
4:35 Eigenanalysis of DTI Tensors Across Populations
Jonathan Taylor, Stanford University

5:00 Random Fields of Multivariate Test Statistics with an Application to Shape Analysis
Keith Worsley, McGill University
5:25 Floor Disscussion

38. RECENT DEVELOPMENTS ON METHODS FOR MULTIVARIATE FAILURE TIME DATA *Salon K*

Sponsors: IMS/ASA Biometrics Section
Organizer: Jianwen Cai, University of North Carolina at Chapel Hill
Chair: Jianwen Cai, University of North Carolina at Chapel Hill

3:45 Recent and Needed Developments in the Analysis of Correlated Failure Time Data
Ross Prentice*, Fred Hutchinson Cancer Research Center and University of Washington
4:10 Some Approaches to Multivariate Failure Time Analysis
Jerry Lawless*, University of Waterloo
4:35 Model Selection for Multivariate Survival Data Analysis
Runze Li*, Penn State University; Jianwen Cai and Haibo Zhou, University of North Carolina at Chapel Hill; Jianqing Fan, Princeton University
5:00 Bayesian Frailty Models Based on Box-Cox Transformed Hazards
Guosheng Yin*, University of Texas M. D. Anderson Cancer Center; Joseph G. Ibrahim, University of North Carolina at Chapel Hill
5:25 Floor Discussion

39. SPECIAL CONTRIBUTED SESSION: ADVANCES IN APPLICATIONS OF LATENT VARIABLE MODEL ANALYSIS FOR HEALTH SERVICES RESEARCH *Salon D*

Sponsor: ENAR
Chair: Karen Bandeen-Roche, Johns Hopkins University

3:45 Optimal Design for Studies with Multivariate Outcomes
Chen-Pin Wang*, University of Texas Health Science Center at San Antonio
4:05 Modeling Heterogeneity in a Randomized Clinical Trial for Treatment of Alcohol Dependence
Jennie Z. Ma* and Chen-Pin Wang, University of Texas Health Science Center at San Antonio
4:25 Causal Inference for Latent Subpopulations
Booil Jo*, Stanford University
4:45 Latent Variable Models for Assessing Treatment Cost-Effectiveness
Alka Indurkhya*, Harvard University
5:05 Mediation Analyses with Structural Mean Models
Tom Tenhave* and Marshall Joffe, University of Pennsylvania School of Medicine
5:25 Floor Discussion

40. CONTRIBUTED PAPERS: CATEGORICAL DATA ANALYSIS AND EXPERIMENTAL DESIGN *Meeting Room 412*

Sponsor: ENAR
Chair: Philip Dixon, Iowa State University

3:45 Nonlinear Test for Categorical Data
Momiao Xiong*, University of Texas Health Science Center at San Antonio
4:00 Evaluation Criteria for Discrete Confidence Intervals: Beyond Coverage and Length
Paul W. Vos and Suzanne S. Hudson*, East Carolina University
4:15 On Hypothesis Testing and Confidence Intervals for the Difference of Two Independent Poisson Rates
William W. B. Wang*, Jin Xu, Santosh Sutradhar, and Ivan S. F. Chan, Merck Research Laboratories
4:30 Multiple Comparisons for Odds Ratios
Melinda H. McCann*, Oklahoma State University; Joshua M. Tebbs, Kansas State University
4:45 Orthogonal Arrays of 2- and 3-Levels for Lean Designs
Changxing Ma*, University of Florida; Ling-Yau Chan, University of Hong Kong
5:00 Power Calculations for a Zero-Inflated Poisson Model
John M. Williamson, Centers for Disease Control and Prevention; Hung-Mo Lin*, Pennsylvania State University College of Medicine; Allen W. Hightower, Centers for Disease Control and Prevention
5:15 Dose-Adjusted Mantel-Haenszel Tests for Numeric Scaled Strata
Stuart A. Gansky*, University of California, San Francisco

41. CONTRIBUTED PAPERS: DESIGNING CLINICAL TRIALS *Salon G*

Sponsor: ENAR
Chair: Beth Ann Griffin, Harvard University

3:45 Adaptive Design and Multiple Comparison Adjustment in Multiple Dose Clinical Trials
Liji Shen*, Sanofi-Aventis
4:00 Sample Size Estimation and Design Selection for a Randomized Trial Subject to Informative Dropouts
Wenjun Li*, University of Massachusetts Medical School
4:15 An Adaptive Dose-Finding Design Incorporating Both Toxicity and Efficacy
Wei Zhang*, University of Iowa; Daniel J. Sargent and Sumithra Mandrekar, Mayo Clinic
4:30 Sample Size Computation for Multivariate Outcomes
Peng Huang*, Barbara C. Tilley, Yuko Palesch, and Jordan Elm, Medical University of South Carolina
4:45 The Value of Information and Optimal Clinical Trial Design
Andrew R. Willan*, Hospital for Sick Children; Eleanor M. Pinto, University of Toronto

5:00 Bivariate Designs in Phase II Trials
Menggang Yu* and Constantin Yiannoutsos, Indiana University School of Medicine

5:15 A Two-Stage Sample Size Recalculation Procedure for Placebo- and Active-Controlled Non-Inferiority Trials
Todd A. Schwartz*, University of North Carolina at Chapel Hill; Jonathan S. Denne, Eli Lilly and Company

42. CONTRIBUTED PAPERS: COMPETING RISKS AND CURE RATES *Salon E*

Sponsor: ENAR
Chair: Guofen Yan, Cleveland Clinic

3:45 A Study of Inverse Probability of Censoring Weighted Estimators of Cumulative Incidence Function for Competing Risks Data
Xu Zhang* and Meijie Zhang, Medical College of Wisconsin

4:00 Nonparametric Estimation with Left Truncated Semi-Competing Risks Data
Limin Peng* and Jason P. Fine, University of Wisconsin–Madison

4:15 Competing Risk Transformation Models with Missingness
Guozhi Gao* and Anastasios A. Tsiatis, North Carolina State University

4:30 Competing Models for Competing Risk
Jack Kalbfleisch and Yining Ye*, University of Michigan

4:45 Flexible Cure Rate Modeling Under Latent Activation Schemes
Freda W. Cooner*, Sudipto Banerjee, and Bradley P. Carlin, University of Minnesota; Debajyoti Sinha, Medical University of South Carolina

5:00 A New Approach to Testing for Sufficient Follow-up in Cure-Rate Analysis
Lev B. Klebanov, Charls University, Praha, Czech Republic; Andrei Y. Yakovlev*, University of Rochester

5:15 Floor Discussion

43. CONTRIBUTED PAPERS: ANALYZING MICROARRAY DATA *Salon C*

Sponsor: ENAR
Chair: Rui Feng, Yale University

3:45 Comparison of Normalization Techniques for cDNA Microarray Data
Kimberly F. Sellers*, University of Pennsylvania; Jeffrey C. Miecznikowski and William F. Eddy, Carnegie Mellon University

4:00 A Non-Parametric Approach of Gene Selection in Oligonucleotide Arrays
Dung-Tsa Chen*, University of Alabama at Birmingham; James Chen, FDA/NCTR; Chen-An Tsai and Seng-jaw Soong, University of Alabama at Birmingham

4:15 **Incorporating Multiple cDNA Microarray Slide Scans—Application to Somatic Embryogenesis in Maize**
Tanzy M. Love* and Alicia L. Carriquiry, Iowa State University

4:30 Statistical Development and Evaluation of Microarray Data Filters
Stanley B. Pounds* and Cheng Cheng, St. Jude Children's Research Hospital

4:45 Linear Mixed Effects Models for Dual Color Microarray Intensity Ratios
Guilherme J. M. Rosa*, Juan Pedro Steibel, and Robert J. Tempelman, Michigan State University

5:00 Probe-Level Correction and Analysis of Affymetrix GeneChips
Fenghai Duan* and Heping Zhang, Yale University School of Medicine

5:15 A Test Statistic for Testing Two-Sample Hypotheses in Microarray Data Analysis
Lev Klebanov, Charls University, Praha, Czech Republic; Alexander Gordon, Yuanhui Xiao*, Hartmut Land, and Andrei Yakovlev, University of Rochester

44. CONTRIBUTED PAPERS: STATISTICAL METHODS IN GENETICS *Meeting Room 410*

Sponsor: ENAR
Chair: Mariza de Andrade, Mayo Clinic College of Medicine

3:45 Approximating Pairwise Alignment Scores with General Assumptions
Lily Wang*, Vanderbilt University; Pranab K. Sen, University of North Carolina at Chapel Hill

4:00 An Improved Multiple Analysis of Associations: A Method for Evaluating the Influence of Multiple Genes on a Trait of Interest
Amy D. Anderson*, Bioinformatics Research Center, North Carolina State University

4:15 Comparison of Methods for the Analysis of Allelic Loss Data
Lei Shen*, Ohio State University

4:30 Genome Phylogenetic Analysis Based on Extended Gene Contents
Xun Gu, Iowa State University; Hongmei Zhang*, University of West Florida

4:45 Likelihood Formulation of Parent-of-Origin Effects for Complex Human Diseases: Characterizing Imprinted Genes for Developmental Dyslexia
Wei Hou*, Cynthia W. Garvan, and Jason G. Craggs, University of Florida; George W. Hynd, Purdue University; Rongling Wu, University of Florida

5:00 Characterizing the Genetic Structure of Populations
Xi Chen* and Bruce S. Weir, North Carolina State University

5:15 Estimating QTL Parameters Under Selective Genotyping
Jaya M. Satagopan*, Memorial Sloan-Kettering Cancer Center; Saunak Sen, University of California, San Francisco; Gary A. Churchill, Jackson Laboratory

Tuesday, March 22
8:30–10:15 a.m.

45. A PRACTICUM ON MULTISTATE SURVIVAL MODELS *Salon A*

Sponsors: ENAR/ASA Biopharmaceutical Section
Organizer: Rick Chappell, University of Wisconsin–Madison
Chair: Rick Chappell, University of Wisconsin–Madison

8:30 A Short Survey of Multistate Models
David Oakes*, University of Rochester Medical Center

9:00 Practical Aspects of Multi-State Models
Terry M. Therneau*, Mayo Clinic

9:30 Models and Statistical Methods for the Current Leukemia Free Survival Function
John P. Klein*, Medical College of Wisconsin

10:00 Floor Discussion

46. STATISTICS IN DISEASE ECOLOGY *Salon B*

Sponsors: ASA Section on Statistics and the Environment/ASA Section on Epidemiology
Organizers: Lance A. Waller, Emory University and G. P. Patil, Pennsylvania State University
Chair: Lance A. Waller, Emory University

8:30 Spatial Contact Networks and Timing of Outbreaks in Epidemic Metapopulations: Theory, Data, and Statistics
Ottar N. Bjornstad*, Pennsylvania State University

9:00 Interval Time Lag Models for Environmental Effects on Arbovirus Positive Mosquito Populations
Frank C. Curriero*, Scott M. Shone, and Greg E. Glass, Johns Hopkins University

9:30 Real-Time Spatial Prediction of Infectious Disease: Experience of New York State (USA) with West Nile Virus and Future Directions for Improved Surveillance
Glen D. Johnson*, New York State Department of Health

10:00 Floor Discussion

47. STATISTICAL METHODS FOR REPRODUCTIVE EPIDEMIOLOGY *Meeting Room 402*

Sponsors: ASA Section on Epidemiology/ASA Section on Survey Research Methods
Organizer: Amy H. Herring, University of North Carolina at Chapel Hill
Chair: David B. Dunson, National Institute of Environmental Health Sciences

8:30 Overview of Methodologic Challenges Facing Reproductive Epidemiology
Germaine M. Louis*, National Institute of Child Health & Human Development

8:55 Statistical Methods for Studying Genetic Contributions to Birth Defects
Clarice R. Weinberg*, David M. Umbach, National Institute of Environmental Health Sciences

9:20 Marginal Regression for Recurrent Marked Point Process Data
Patrick J. Heagerty*, University of Washington

9:45 The Design of the National Children's Study: Probability Sample, Medical Center Model, or What?
Roderick J. A. Little*, University of Michigan

10:10 Floor Discussion

48. MODELING BRAIN IMAGES—THE EFFECTS OF SPACE, TIME, AND INDIVIDUALITY *Salon K*

Sponsor: ENAR
Organizer: William F. Eddy, Carnegie Mellon University
Chair: William F. Eddy, Carnegie Mellon University

8:30 Revealing Brain Activity with Filters
Kary L. Myers*, Carnegie Mellon University

9:00 Spatio-Temporal Modeling of Localized Brain Activity
DuBois Bowman*, Emory University

9:30 Robust and Local Nonsphericity Modeling for Second-Level PET and fMRI Analysis
Thomas E. Nichols*, Jeanette Mumford, and Wen-Lin Luo, University of Michigan

10:00 Floor Discussion

49. CROSSOVER DESIGNS FOR THE PHARMACEUTICAL INDUSTRY *Salon G*

Sponsor: IMS
Organizer: Min Yang, University of Nebraska-Lincoln
Chair: Min Yang, University of Nebraska-Lincoln

8:30 Optimal and Efficient Crossover Designs When Subject Effects Are Random
John Stufken*, University of Georgia
8:55 Efficient Designs for Experiments with Multiple Treatments per Subject
James L. Rosenberger*, Pennsylvania State University
9:20 Crossover Designs for Comparing Test Treatments with a Control
Sam Hedayat*, University of Illinois at Chicago
9:45 On Optimal Cross-Over Designs When Carry-Over Effects Are Proportional to Direct Effects
R. A. Bailey, Queen Mary, University of London; J. Kunert*, Universität Dortmund
10:10 Floor Discussion

50. CONTRIBUTED PAPERS: DIAGNOSTIC TESTS *Meeting Room 415*

Sponsor: ENAR
Chair: Liji Shen, Sanofi-Aventis

8:30 The Predictive Distribution and Diagnostic Accuracy
Lyle D. Broemeling* and Marcella Johnson, University of Texas M. D. Anderson Cancer Center
8:45 Peak-Picking Algorithm for LC-ESI-FT Mass Spectrometry Data
Jeanette E. Eckel-Passow*, Ann L. Oberg, Terry M. Therneau, Chris J. Mason, and David C. Muddiman, Mayo Clinic College of Medicine
9:00 A Novel Algorithm for MALDI-TOF MS Data Processing Using Mathematical Tools
Shuo Chen*, East Tennessee State University; Don Hong and Yu Shyr, Vanderbilt University
9:15 Resolving the Degrees of Freedom Issue Concerning the Dorfman-Berbaum-Metz and Obuchowski-Rockette Methods for Receiver Operating Characteristic (ROC) Data
Stephen L. Hillis*, Iowa City V.A. Medical Center
9:30 A Framework for the Study of the Predictive Accuracy of Diagnostic Tests
Shang-Ying Shiu* and Constantine Gatsonis, Brown University
9:45 A Permutation Test Sensitive to Differences in Areas for Comparing ROC Curves from a Paired Design
Andriy I. Bandos*, Howard E. Rockette, and David Gur, University of Pittsburgh
10:00 Floor Discussion

51. CONTRIBUTED PAPERS: MIXED MODELS: LINEAR, GENERALIZED, AND NON-LINEAR *Meeting Room 410*

Sponsor: ENAR
Chair: Haiying Chen, Wake Forest University

8:30 Linear Mixed Models with Skewed t Distributions
Tianyue Zhou*, University of Illinois at Urbana-Champaign
8:45 Analytical Methods for Compliance with Longitudinal Assessments in Clinical Studies
Stephanie R. Land* and Ritter Marcie, University of Pittsburgh
9:00 A Diagnostic Test for the Mixing Distribution in a Generalized Linear Mixed Model
Eric J Tchetgen* and Brent Coull, Harvard University
9:15 Bayesian Covariance Selection in Generalized Linear Mixed Models
Bo Cai* and David B. Dunson, National Institute of Environmental Health Sciences
9:30 Small Sample Inference for Cluster Samples with Gaussian Data
Jacqueline L. Johnson*, Diane J. Catellier, and Keith E. Muller, University of North Carolina at Chapel Hill
9:45 Modeling Inter-Rater Agreement Using Mixed Models
Kerrie Nelson* and Don Edwards, University of South Carolina
10:00 A Non-Linear Mixed Effect Model for Hepatitis C Viral Dynamics
Abdus S. Wahed*and Kyungah Im, University of Pittsburgh; Thelma Wiley, Rush University; Steven Belle, University of Pittsburgh

52. CONTRIBUTED PAPERS: BAYESIAN METHODS *Salon J*

Sponsor: ENAR
Chair: Dawei Xie, University of Pennsylvania

8:30 Bayesian Inference in Generalized Additive Mixed Models with Nonparametric Random Effects
Yisheng Li*, University of Texas M. D. Anderson Cancer Center; Xihong Lin, University of Michigan
8:45 Hierarchical Models for a Time Series on Marijuana Abuse Among Hospital Emergency Room Admissions
Li Zhu*, Dennis Gorman, Scott Horel, and Wen Tan, Texas A&M University
9:00 **Bayesian Adaptive Regression Splines for Hierarchical Data**
Jamie L. Bigelow* and David B. Dunson, National Institute of Environmental Health Sciences
9:15 Bayesian Free Knot Curve Fitting with Applications to Spectral Density Estimation
Carsten H. Botts*, University of Florida and Iowa State University; Michael Daniels, University of Florida

9:30 Prediction of Protein Inter-Domain Linker Regions by a Hidden Markov Model
Kyounghwa Bae*, Christine G. Elsik, and Bani K. Mallick, Texas A&M University

9:45 Reparameterization to Improve Bayesian Computing for the CAR Model with Two Types of Neighbor Relations
Yi He* and James S. Hodges, University of Minnesota

10:00 Clustering Analysis of Ordinal Data
Xian Zhou*, Peter Mueller, and Nebiyou Bekele, University of Texas M. D. Anderson Cancer Center

53. CONTRIBUTED PAPERS: MISSING DATA METHODS
Salon D

Sponsor: ENAR
Chair: Monnie McGee, Southern Methodist University

8:30 Challenges of Non-Ignorable Missing Data in Clinical Trials: A Pattern Mixture Model Approach
G. K. Balasubramani*, Stephen R. Wisniewski, and James Luther, University of Pittsburgh

8:45 Regression Analyses with Data Missing at Random—An Extension of the EM Algorithm
Yang Y. Zhao*, Jerald F. Lawless, and Donald L. McLeish, University of Waterloo

9:00 A Hierarchical Technique for Estimating Location Parameter in the Presence of Missing Data
Sergey S. Tarima* , University of Kentucky ;Yuriy G. Dmitriev, Tomsk State University, Russia; Richard J. Kryscio, University of Kentucky

9:15 Effects of Methods of Estimation of Missing Data
J. Lynn Palmer*, University of Texas M. D. Anderson Cancer Center

9:30 Correlating Two Continuous Variables Subject to Detection Limits in the Context of Mixture Distributions
Haitao Chu* and Lawrence H. Moulton, Johns Hopkins University; Wendy J. Mack, University of Southern California; Douglas J. Passaro, University of Illinois at Chicago; Paulo F. Barroso, Hospital Universitário Clementino Fraga Filho School of Medicine, Universidade Federal do Rio de Janeiro, Brazil; Alvaro Muñoz, Johns Hopkins Bloomberg School of Public Health

9:45 Sample Size Calculation for Longitudinal Data Under Missing at Random (MAR)
Susan Halabi*, Daohai Yu, and Sin-Ho Jung, Duke University Medical Center

10:00 Floor Discussion

54. CONTRIBUTED PAPERS: METHODS FOR MULTIPLE ENDPOINTS
Salon E

Sponsor: ENAR
Chair: Adin-Cristian Andrei, University of Michigan

8:30 Multivariate Analysis of Binary Data from Drug Safety Trials
Bernhard Klingenberg*, Williams College; Alan Agresti, University of Florida

8:45 A Marginalized Diffusion Model for Combining State and National Level Survey Data
Diana L. Miglioretti*, Group Health Cooperative; Elizabeth Brown, University of Washington

9:00 Modeling Differentiated Treatment Effects for Multiple Outcomes Data
Hongfei Guo*, Johns Hopkins University

9:15 Avoiding Test Size Bias Due to Internal Pilots with Gaussian Repeated Measures
Meagan E. Clement* , University of North Carolina at Chapel Hill; Christopher S. Coffey, University of Alabama at Birmingham ; Keith E. Muller, University of North Carolina at Chapel Hill

9:30 Comparing Treatment Means in a Repeated Measures Analgesic Study
Guoyong Jiang* and Lilliam Kingsbury, Cephalon, Inc.

9:45 Mixed-Effects Probit Model for Longitudinal Data with Multiple Discrete Outcomes
Robert Gibbons, Hua Yun Chen*, and Dullal Bauhmik, University of Illinois at Chicago

10:00 Marginalized Regression Models for Long Series of Longitudinal Binary Response Data
Jonathan S. Schildcrout*, Vanderbilt University; Patrick J. Heagerty, University of Washington

55. CONTRIBUTED PAPERS: QUANTITATIVE-TRAIT LINKAGE ANALYSIS *Meeting Room 412*

Sponsor: ENAR
Chair: Hoa Nguyen, Carnegie Mellon University.

8:30 A Joint Model for Nonparametric Functional Mapping of Growth Curves and Time-to-Events
Min Lin* and Rongling Wu, University of Florida

8:45 Variable Selection for Large P Small n Regression Model with Incomplete Data: Application to QTL Mapping
Min Zhang*, Cornell University; Dabao Zhang, University of Rochester Medical Center; Martin T. Wells, Cornell University

9:00 Robust Semiparametric Multipoint Quantitative-Trait Linkage Analysis in General Pedigrees
Guoqing Diao* and Danyu Lin, University of North Carolina at Chapel Hill

9:15 Structured Antedependence Models
Wei Zhao*, Wei Hou, Ramon Littell, and Rongling Wu, University of Florida

9:30 Using Generalized Estimating Equations in a Genome Scan of Cell Counts in the Dentate Gyrus of Recombinant Inbred Mice
Dirk F. Moore*, University of Medicine and Dentistry of New Jersey

9:45 A Statistical Framework for Functional Mapping of Intracellular Circadian Rhythms
Tian Liu*, Fei Long, and Rongling Wu, University of Florida

10:00 Floor Discussion

Tuesday, March 22
10:15–10:30 a.m.

Break *Grand Ballroom Pre-Function Area*

Tuesday, March 22
10:30 a.m.–12:15 p.m.

PRESIDENTIAL INVITED ADDRESS *Salon H*

Sponsor: ENAR
Organizer/Chair: Peter B. Imrey, Cleveland Clinic Foundation

10:30 Introduction: Peter B. Imrey, Cleveland Clinic Foundation

10:40 Distinguished Student Paper Awards:
Timothy G. Gregoire, Yale University

10:50 Selection and Estimation for Large-Scale Simultaneous Inference
Bradley Efron, Stanford University

Tuesday, March 22
1:45–3:30 p.m.

56. SCREENING FOR DISEASE: ISSUES IN STUDY DESIGN AND ANALYSIS *Salon A*

Sponsors: ASA Section on Survey Research Methods/ASA Biopharmaceutical Section
Organizer: Nancy Obuchowski, Cleveland Clinic Foundation
Chair: Steve Hillis, University of Iowa

1:45 Estimating the Accuracy of CT Colonography
Valerie L. Durkalski*, Yuko Palesch, and Peter B. Cotton, Medical University of South Carolina

2:15 Challenges to Studies of Screening Tests with Illustrations from Total Body Screening with CT
Nancy A. Obuchowski*, Cleveland Clinic Foundation

2:45 ROC Curve Evaluation of Screening Modalities
Constantine Gatsonis*, Brown University; Mei Hsiu Chen, University of California, San Francisco

3:15 Discussant: Colin Begg, Memorial Sloan-Kettering Cancer Center

57. COMBINING INFORMATION ACROSS SPATIAL SCALES *Salon B*

Sponsors: ASA Section on Statistics and the Environment/ASA Section on Statistics in Defense and National Security
Organizer: Philip Dixon, Iowa State University
Chair: Philip Dixon, Iowa State University

1:45 Modeling Global Covariance Structures Using Local Information
Petrutza C. Caragea*, Iowa State University
2:15 Modeling Nonhomogeneous Poisson Processes Using a Combination of Point and Count Data
Mark S. Kaiser* and Han Wu, Iowa State University
2:45 Combining Data Sources to Evaluate Spatial and Temporal Patterns of Avian Population Change
William A. Link* and John R. Sauer, USGS Patuxent Wildlife Research Center
3:15 Discussant: Mary Christman, University of Florida

58. BAYESIAN PROCEDURES FOR ANALYZING MICROARRAY DATA *Salon G*

Sponsor: ENAR
Organizer: E. Olusegun George, Memphis State University
Chair: E. Olusegun George, Memphis State University

1:45 Comparative Genomics Analysis of Gene Regulation
Jun Liu*, Cristian Castillo-Davis, and Lei Shen, Harvard University
2:15 Towards a Complete Picture of Gene Regulation: Using Bayesian Approaches to Integrate Genomic Sequence and Expression Data
Mayetri Gupta* and Joseph G. Ibrahim, University of North Carolina at Chapel Hill
2:45 Bayesian Variable Selection Methods for the Analysis of DNA Microarray Data
Mahlet G. Tadesse*, University of Pennsylvania; Naijun Sha, University of Texas at El Paso; Marina Vannucci, Texas A&M University
3:15 Floor Discussion

59. COST-EFFECTIVENESS ANALYSIS: METHODOLOGIES FOR COMPARING COMPETING HEALTH CARE INTERVENTIONS (THEME SESSION) *Salon D*

Sponsor: ASA Health Policy Statistics Section
Organizer: Joseph Gardiner, Michigan State University
Chair: Joseph Gardiner, Michigan State University

1:45 Designing National Health Care Expenditure Surveys to Inform Health Policy and Practice
Steven B. Cohen*, AHRQ
2:15 QoL Adjustments for Nondegradation Processes in Life Time Analysis
Pranab K. Sen*, University of North Carolina at Chapel Hill
2:45 A Dynamic Model to Assess Covariate Effects on Cost and Health Outcomes
Joseph C. Gardiner*, Zhehui Luo, Corina M. Sirbu, Cathy J. Bradley, and Charles W. Given, Michigan State University
3:15 Floor Discussion

60. RECENT ADVANCES IN SEMIPARAMETRIC ESTIMATION *Meeting Room 402*

Sponsors: ENAR/IMS
Organizer: David Ruppert, Cornell University
Chair: David Ruppert, Cornell University

1:45 Semiparametric Spatial Modeling of Binary Outcomes
Tatiyana V. Apanasovich* and David Ruppert, Cornell University; Raymond J. Carroll, Texas A&M University
2:10 Transfer Functions in Hierarchical Functional Data Models, with Applications to Predicting Cell Proliferation and Apoptosis from p27 Expression in Colon Carcinogenesis Experiments
Raymond J. Carroll*, Veera Baladandayuthapani, and Kimberly Drews, Texas A&M University
2:35 Spatially Adaptive Bayesian P-Splines with Heteroscedastic Errors
Ciprian M. Crainiceanu*, Johns Hopkins University
3:00 A Semiparametric Model for a Nonstationary Time Series of Counts with Time-Dependent Covariates
Andres Houseman, Brent A. Coull*, and James P. Shine, Harvard University
3:25 Floor Discussion

61. CONTRIBUTED PAPERS: BIOASSAY AND BIOPHARMACEUTICAL APPLICATIONS

Meeting Room 415

Sponsor: ENAR
Chair: Russell Stocker, Mississippi State University

1:45 Segment Response Surface for Synergy Analysis
Xiaoli S. Hou* and Keith A. Soper, Merck & Co., Inc.

2:00 Testing Immunological Correlates of Protection
Andrew J. Dunning*, Wyeth Vaccines Research

2:15 Assessing In Vitro Bioequivalence for Profile Data: A New Modeling Approach
Bin Cheng*, Columbia University

2:30 Reducing Animal Use in Chemical Toxicity Testing
Robert Lee*, Eric Harvey, and Patrick Crockett, Constella Health Sciences; Shyamal Peddada, National Institute of Environmental Health Sciences

2:45 Statistical Analysis of CDFSE Data
Ollivier Hyrien*, University of Rochester Medical Center

3:00 A Response Surface Model for Drug Combinations
Maiying Kong* and J. Jack Lee, University of Texas M. D. Anderson Cancer Center

3:15 Comparing Synergy
Gregory D. Ayers* and J. Jack Lee, University of Texas M. D. Anderson Cancer Center

62. CONTRIBUTED PAPERS: SEQUENTIAL METHODS

Salon E

Sponsor: ENAR
Chair: Mimi Kim, Albert Einstein College of Medicine of Yeshiva University

1:45 A Sequential Test for Treatment Effects Under Staggered Entry in a Multicenter Clinical Trial
Dong-Yun Kim*, Illinois State University; Michael B. Woodroofe, University of Michigan

2:00 Supplementary Analysis of Probabilities at the Termination of a Group Sequential Phase II Trial
Aiyi Liu*, Chengqing Wu, and Kai F. Yu, National Institute of Child Health and Human Development; Edmund A. Gehan, Georgetown University Medical Center

2:15 Exact Group Sequential Designs for Detecting Hypotheses Involving Two Correlated Binary Responses
Jihnhee Yu*, James L. Kepner, and Brian N. Bundy, Roswell Park Cancer Institute

2:30 Combinations of Two-Stage Designs for Testing Multiple Treatments in Phase II Cancer Trials
Tatsuki Koyama*, Jordan D. Berlin, and Yu Shyr, Vanderbilt University School of Medicine

2:45 Bayesian Optimal Design for Phase II Trials
Meichun Ding*, Rice University; Gary L. Rosner and Peter Mueller, University of Texas M. D. Anderson Cancer Center

3:00 Conditional Maximum Likelihood Estimation Following a Group Sequential Test
Aiyi Liu, James Troendle, and Kai F. Yu, DHHS/NIH/NICHD/DESPR; Weishi Yuan*, DHHS/FDA

3:15 Estimating Rates of Response and Toxicity Following a Bivariate Group Sequential Phase II Clinical Trial
Chengqing Wu*, Aiyi Liu, and Kai F. Yu, BMSB,DESPR, NICHD, NIH; Ming Tan, Greenebaum Cancer Center, University of Maryland

63. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS II

Meeting Room 410

Sponsor: ENAR
Chair: Yisheng Li, University of Texas M. D. Anderson Cancer Center

1:45 Semiparametric Joint Modeling of Longitudinal Measurements and Time-to-Event Data
Wen Ye*, Xihong Lin, and Jeremy M. G. Taylor, University of Michigan

2:00 **Estimation in Two-Stage Randomization Designs**
Xiang Guo* and Anastasios A. Tsiatis, North Carolina State University

2:15 Nonparametric Regression Using Kernel Estimating Equations for Correlated Failure Time Data
Zhangsheng Yu* and Xihong Lin, University of Michigan

2:30 Methods for Analyzing Survival Data from Alternating Studies
Beth Ann Griffin* and Stephen Lagakos, Harvard University

2:45 A Sequential Stratification Method to Estimate the Effect of a Time-Dependent Treatment in the Analysis of Recurrent Event Data
Douglas E. Schaubel* and Robert A. Wolfe, University of Michigan

3:00 Semiparametric Transformation Models for the Case-Cohort Study
Wenbin Lu* and Anastasios Tsiatis, North Carolina State University

3:15 Floor Discussion

64. CONTRIBUTED PAPERS: LONGITUDINAL DATA ANALYSIS AND GENERALIZED LINEAR MODELS *Meeting Room 412*

Sponsor: ENAR
Chair: Justine Shults, University of Pennsylvania

1:45 Matrix Skewed Distributions
Solomon W. Harrar*, South Dakota State University; Arjun K. Gupta, Bowling Green State University

2:00 Longitudinal Data Analysis in F1-LD-F1 Factorial Design with Application to Liver Cancer Study
Ke Yan*, Texas Tech University

2:15 Methods of Longitudinal Data for F2-LD-F1 Model
Lan Zhang*, Texas Tech University

2:30 Controlling Variable Selection By the Addition of Pseudo-Variables
Yujun Wu*, University of Medicine and Dentistry of New Jersey; Dennis D. Boos and Leonard A. Stefanski, North Carolina State University

2:45 Learning Curve Analysis of Logistic Regression and Tree-Structured Algorithms
Qinghua Song* and Wei-Yin Loh, University of Wisconsin–Madison; Kin Yee Chan, National University of Singapore, Republic of Singapore; Yu-Shan Shih, National Chung Cheng University, Taiwan, Republic of China

3:00 Transformation Supporting EDF Goodness-of-Fit Test of Normal Distribution Related to Two Samples Based on Regression
Dhanuja Kasturiratna*, Truc T. Nguyen, and Arjun K. Gupta, Bowling Green State University

3:15 The Role of Percentiles in Determining a Regression Equation
Yvonne M. Zubovic* and Chand K. Chauhan, Indiana University Purdue University Fort Wayne

65. CONTRIBUTED PAPERS: BAYESIAN METHODS IN CLINICAL TRIALS *Salon J*

Sponsor: ENAR
Chair: J. Lynn Palmer, University of Texas M. D. Anderson Cancer Center

1:45 Proof of Concept Trials
A. Lawrence Gould*, Merck Research Laboratories

2:00 Bayesian Sample Size Calculations in Phase II Clinical Trials Using a Simple Mixture of Informative Priors: Incorporating Pessimistic and Optimistic Opinions Simultaneously
Byron J. Gajewski* and Matthew S. Mayo, University of Kansas Medical Center

2:15 Bayesian Estimation of Cost-Effectiveness Using Pattern-Mixture Models
Clara Y. Kim* and Daniel F. Heitjan, University of Pennsylvania

2:30 A Cardiology Trial Optimal Design Constrained by Ethical Considerations
Manuela Buzoianu* and Joseph B. Kadane, Carnegie Mellon University

2:45 Do Antidepressants Cause Suicide in Children? A Bayesian Meta-Analysis
Eloise E. Kaizar*, Joel Greenhouse, and Howard Seltman, Carnegie Mellon University

3:00 Floor Discussion

66. CONTRIBUTED PAPERS: BAYESIAN METHODS IN GENOMIC DATA ANALYSIS *Salon K*

Sponsor: ENAR
Chair: Bonnie LaFleur, Vanderbilt University

1:45 Exploratory Bayesian Model Selection for High-Order SNP-Phenotype Associations
Jing X. Zhao*, Merck Research Laboratories; Andrea S. Foulkes, University of Massachusetts; Edward I. George, Muredach Reilly, and Daniel J. Rader, University of Pennsylvania

2:00 Normalization of Microarrays in Transcription Inhibition
Yan Zheng* and Cavan Reilly, University of Minnesota

2:15 **A Bayesian Method for Finding Interactions in Genomic Studies**
Wei Chen*, Debashis Ghosh, Trivellore Raghunathan, and Sharon Kardia, University of Michigan

2:30 Estimating Model Complexity for Bayesian Network Learning
Anthony Almudevar and Peter Salzman*, University of Rochester

2:45 Incorporating Prior Information via Shrinkage: A Combined Analysis of Genome-Wide Location Data and Gene Expression Data
Yang Xie*, Wei Pan, Keyong S. Jeong, and Arkady Khodursky, University of Minnesota

3:00 Empirical Bayes Inference for Partially Nested Designs in Two-Color Microarray Systems
Robert J. Tempelman*, Michigan State University

3:15 Floor Discussion

Tuesday, March 22
3:30–3:45 p.m

Break *Grand Ballroom Pre-Function Area*

Tuesday, March 22
3:45–5:30 p.m.

67. METHODOLOGIC ISSUES IN STUDIES INVOLVING CANCER SCREENING *Salon A*

Sponsors: ASA Biopharmaceutical Section/ASA Section on Epidemiology
Organizer: Marshall Joffe, University of Pennsylvania
Chair: Marshall Joffe, University of Pennsylvania

3:45 Causal Inference and Screening
James M. Robins*, Harvard University
4:15 Bias from Variability in Diagnostic Delay
Timothy R. Church*, University of Minnesota
4:45 A Novel Alternative to Cause-Specific Mortality for Evaluating Cancer Screening
Marshall M. Joffe*, University of Pennsylvania
5:15 Floor Discussion

68. CURRENT ADVANCES IN MODELING TIME COURSE GENE EXPRESSION DATA *Salon G*

Sponsors: ENAR/IMS
Organizer: Naisyin Wang, Texas A&M University
Chair: Naisyin Wang, Texas A&M University

3:45 Hidden Markov Models for Microarray Time Course and Cell Cycle Data in Multiple Biological Conditions
Christina Kendziorski* and Ping Wang, University of Wisconsin–Madison; Ming Yuan, Georgia Institute of Technology
4:15 Statistical Methods for Analysis of Microarray Time Course Gene Expression Data
Hongzhe Li*, University of California, Davis; Fangxin Hong, Salk Institute
4:45 A Sequential Bayesian Approach with Applications to Circadian Rhythm Microarray Gene Expression Data
Faming Liang*, Chuanhai Liu and Naisyin Wang, Texas A&M University
5:15 Floor Discussion

69. STATISTICAL ISSUES IN A BIOMETRIC SURVEY: THE NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES) *Salon B*

Sponsors: ASA Section on Survey Research Methods
Organizer: Lester R. Curtin, Centers for Disease Control and Prevention
Chair: Lester R. Curtin, Centers for Disease Control and Prevention

3:45 Indicators of Obesity from NHANES
Allison Hedley* and Cynthia Ogden, Centers for Disease Control and Prevention
4:15 Statistical Issues in Analyzing Environmental Health Data from the NHANES
Susan Schober*, Lisa Mirel, and Lester Curtin, Centers for Disease Control and Prevention
4:45 Studies of Cardiovascular Fitness in NHANES
Chia-Yih Wang*, Jeffrey Hughes, and Lester Curtin, Centers for Disease Control and Prevention
5:15 Floor Discussion

70. STRUCTURAL EQUATIONS AND PSYCHOMETRIC METHODS IN BIOLOGICAL STUDIES (THEME SESSION) *Salon D*

Sponsors: ENAR/ASA Biometrics Section
Organizer: David B. Dunson, National Institute of Environmental Health Sciences
Chair: Bo Cai, National Institute of Environmental Health Sciences

3:45 Structural Models with Nonnormal, Missing, Multilevel, or Atypical Data: The EQS Approach
Peter M. Bentler*, University of California, Los Angeles
4:10 Sparse Regressions and Graphical Models
Mike West*, Adrian Dobra, Chris Hans, and Carlos Carvalho, Duke University
4:35 Bayesian Latent Variable Density Regression
David B. Dunson*, National Institute of Environmental Health Sciences
5:00 Mixed-Effects Variance Components Models for Biometric Family and Longitudinal Analyses
John J. McArdle*, University of Virginia; Carol A. Prescott, Virginia Commonwealth University
5:25 Floor Discussion

71. SEEKING PATTERN IN GENOMIC DATA USING CROSS-SPECIES COMPARISONS *Salon K*

Sponsor:IMS
Organizer: Naomi S. Altman, Pennsylvania State University
Chair: Annie Qu, Oregon State University

3:45 Global Classification of (Plant) Proteins Across Multiple Species
Kerr Wall, Jim Leebens-Mack, Naomi S. Altman*, and Claude dePamphilis, Pennsylvania State University; Victor Albert, Natural History Museums and Botanical Garden, University of Oslo; Dawn Field, Oxford University

4:15 Studying Functional Non-Coding Sequences Through Supervised and Unsupervised Analyses of Genomic Alignment Data
James Taylor, Webb Miller, and Francesca Chiaromonte*, Pennsylvania State University

4:45 Investigation of Plant Phosphorylation Using Intergenomic Comparison
Michael Gribskov*, Purdue University

5:15 Floor Discussion

72. CONTRIBUTED PAPERS: SEMIPARAMETRIC AND NONPARAMETRIC MODELING *Meeting Room 412*

Sponsor: ENAR
Chair: Susanne May, University of California, San Diego

3:45 Confidence Intervals for Semi-Parametric Quantile Regression
Mi-Ok Kim*, University of Kentucky

4:00 Capturing Higher-Order Features in Pooling Strategy with Kernel Logistic Model
Peter X. K. Song and Peng Zhang*, University of Waterloo; Rui Liu, York University

4:15 **Semiparametric Regression for High-Dimensional Data with Applications in Microarrays: Least Square Kernel Machines and Linear Mixed Models**
Dawei Liu*, Xihong Lin, and Debashis Ghosh, University of Michigan

4:30 On a Flexible Information Criterion for Order Selection in Finite Mixture Models
Richard Charnigo*, University of Kentucky; Ramani S. Pilla, Case Western Reserve University

4:45 Testing Lack-of-Fit of Nonlinear Regression Models via Local Linear Regression Techniques
Chin-Shang Li*, St. Jude Children's Research Hospital

5:00 Nonparametric Spline Estimators of Comparison Distribution Functions and ROC Curves
Piea Peng Lee* and Andrzej Kozek, Macquarie University, Sydney, Australia

5:15 Improving Regression Function Estimators
Ali Khoujmane*, Texas Tech University

73. CONTRIBUTED PAPERS: IMAGING OF BRAIN ACTIVITY *Meeting Room 415*

Sponsor: ENAR
Chair: Chun Li, Vanderbilt University

3:45 Increasing the Power of Group Comparisons in SPECT Brain Imaging Through Spatial Modeling of Intervoxel Correlations
Jeffrey S. Spence*, Patrick S. Carmack, and Robert W. Haley, University of Texas Southwestern Medical Center; Richard F. Gunst, William R. Schucany and Wayne A. Woodward, Southern Methodist University

4:00 A Hierarchical Deformation Model for Images
Sining Chen*, Sidney Kimmel Comprehensive Cancer Center; Helene Benveniste, Brookhaven National Laboratory; Valen E. Johnson, University of Texas M. D. Anderson Cancer Center

4:15 A Bayesian Approach to Determining Connectivity of the Human Brain
Rajan S. Patel* and F. Dubois Bowman, Emory University

4:30 A Mixture Approach for PET Studies
Huiping Jiang*, New York State Psychiatric Institute; Todd Ogden, Columbia University

4:45 3D Wavelet Denoising of SPECT Images
Liansheng Tang*, William R. Schucany, and Wayne A. Woodward, Southern Methodist University

5:00 The Brain as a Mediator: Where Does This Affect That?
David M. Shera*, Lijun Jing, and Tony J. Simon, University of Pennsylvania

5:15 Floor Discussion

74. CONTRIBUTED PAPERS: COX REGRESSION MODELS
Salon E

Sponsors: ENAR
Chair: Byron Gajewski, University of Kansas Medical Center

3:45 Bayesian Variable Selection in Cox Regression Models
Naijun Sha*, University of Texas at El Paso; Mahlet Tadesse, University of Pennsylvania; Marina Vannucci, Texas A&M University

4:00 General Instrumental Variables Estimation in Cox's Proportional Hazards Model with Time-Varying Treatment
David S. Cohen*, Thomas A. Louis, and Daniel O.Scharfstein, Johns Hopkins University; Sam Bozzette, Henry K. Tam, and Christopher F. Ake, Veterans Affairs San Diego Healthcare System

4:15 Assessment of the Cox Model for Binary Interval-Censored Failure Time Data
Lianming Wang*, University of Missouri

4:30 Incorporating Time-Dependent Covariates in Survival Analysis Using the LVAR Method
Yali Liu* and Bruce A. Craig, Purdue University

4:45 A Comparison of Statistical Tests for Assessing the Proportional Hazards Assumption in the Cox Model
Inger Persson, Trial Form Support, AB, Stockholm, Sweden; Harry J. Khamis*, Wright State University

5:00 Examining Model Fit for Exposure-Response Curves with Penalized Splines in Cox Models
Elizabeth J. Malloy* and Ellen A. Eisen, Harvard University

5:15 Testing the Proportional Odds Model for Interval-Censored Survival Data
Jianguo Sun, University of Missouri–Columbia; Liuquan Sun, Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing; Chao Zhu*, University of Missouri–Columbia

75. CONTRIBUTED PAPERS: LONGITUDINAL DATA ANALYSIS
Meeting Room 410

Sponsor: ENAR
Chair: John Jiang, Cephalon, Inc.

3:45 Marginalized Transition Models for Longitudinal Polytomous Data
Keunbaik Lee* and Michael Daniels, University of Florida

4:00 Joint Analysis of Longitudinal Data with Informative Right Censoring
Mengling Liu*, New York University School of Medicine; Zhiliang Ying, Columbia University

4:15 Using Trajectories from a Bivariate Growth Curve of Covariates in a Cox Model Analysis
Qianyu Dang*, Sati Mazumdar, and Stewart J. Anderson, University of Pittsburgh

4:30 Functional Data Analysis Issues for Identifying Placebo Response in Drug Treated Subjects
Thaddeus Tarpey*, Wright State University; Eva Petkova and Todd Ogden, Columbia University

4:45 Estimating Health Outcomes Trajectories via Finite Mixture Models
Jason T. Connor*, Carnegie Mellon University; Susana Arrigain, Cleveland Clinic Foundation

5:00 A Likelihood-Based Approach to the Analysis of Colonic Crypt Signaling
Kimberly L. Drews* and Raymond J. Carroll, Texas A&M University

5:15 Floor Discussion

76. CONTRIBUTED PAPERS: TOPICS IN BIOSTATISTICS: FROM QUANTAL RESPONSES TO CASE-CONTROL STUDIES
Meeting Room 402

Sponsor: ENAR
Chair: Bin Cheng, Columbia University

3:45 Case-Control Studies with Longitudinal Covariates
Honghong Zhou*, Xihong Lin, and Bin Nan, University of Michigan

4:00 A Strategy for Assessing the Performance of Alternatives to CD4 Cell Count in Making the Decision to Start Antiretroviral Treatment in Resource Constrained Settings
Lisa M. Wruck*, New England Research Institutes; Michael D. Hughes, Harvard University

4:15 Sample Size Determination with FDR Adjustment for Microarray Experiments
Gengqian Cai* and Sanat K. Sarkar, Temple University

4:30 Multivariable Binary Regression with Stochastic Covariates
Evrim Oral*, Middle East Technical University, Ankara,Turkey; Diana L. Miglioretti, Center for Health Studies, Group Health Cooperative

4:45 Comparison of Logistic Regression Versus Propensity Scoring in Binary Treatment Effect Estimation
Yi Huang*, Karen Bandeen-Roche, and Constantine Frangakis, Johns Hopkins Unversity

5:00 A Note on Tests for Interaction in Quantal Response Data
Melinda M. Holt*, Southeastern Louisiana University; James Stamey, Stephen F. Austin University; John W. Seaman Jr., and Dean M. Young, Baylor University

5:15 Agreement for Curved Data
Jason Liao*, Merck Research Laboratory

77. CONTRIBUTED PAPERS: FROM PROTEOMICS TO CLASSIFICATION TO MEASUREMENT ERROR: CURRENT TOPICS IN PROTEOMICS/GENOMICS *Salon J*

Sponsor: ENAR
Chair: Don Hong,Vanderbilt, University/ETSU

3:45 A Method for Spot Finding in Two-Dimensional Images
Jeffrey C. Miecznikowski*, William F. Eddy, and Jonathan S. Minden, Carnegie Mellon University; Kimberly F. Sellers, University of Pennsylvania

4:00 A Conditional Moment Method for Model Selection in Penalized Logistic Regression for Disease Classification Using Microarray Gene Expression Data
J. G. Liao*, University of Medicine and Dentistry, New Jersey

4:15 Gene Function Prediction by a Combined Analysis of Gene Expression Data and Protein-Protein Interaction Data
Guanghua Xiao* and Wei Pan, University of Minnesota

4:30 CLASSIX: A New Classification Method Based on a Separation Index
Weiliang Qiu* and Mei-Ling T. Lee, Channing Laboratory–Brigham and Women's Hospital, Harvard Medical School

4:45 Structural Equation Modeling of Genotype by Environment Interaction
Prabhakar Dhungana*, Kent M. Eskridge, P. S. Baenziger, Lenis Nelson, W. Stroup, and Albert Weiss, University of Nebraska; B. T. Campbell, USDA

5:00 Measurement Error Model for cDNA Microarray and Time-to-Event Data
Jonathan A. L. Gelfond* and Joseph G. Ibrahim, University of North Carolina at Chapel Hill

5:15 Floor Discussion

Wednesday, March 23
8:30–10:15 a.m.

78. ADVANCED TOPICS IN PROSTATE CANCER MODELING: A MULTIDISCIPLINARY AND INTEGRATED PERSPECTIVE *Salon A*

Sponsors: ASA Biopharmaceutical Section/ASA Section on Epidemiology
Organizer: Kim-Anh Do, University of Texas M. D. Anderson Cancer Center
Chair: Kim-Anh Do,University of Texas M. D. Anderson Cancer Center

8:30 Statistical Modeling Strategies to Define the Biologic Basis of Clinically Significant Prostate Cancer
Timothy J. McDonnell*, University of Texas M. D. Anderson Cancer Center

8:55 Correlating Microarray Gene Expression Measurements with Gleason Scores and Identifying Biomarkers to Distinguish Prostate Cancer Stages
Jing Wang*, Kim-Anh Do, Sijin Wen, Spyros Tsavachidis, Timothy J. McDonnell, and Kevin R. Coombes, University of Texas M. D. Anderson Cancer Center

9:20 Combining Longitudinal Studies of PSA
Lurdes YT Inoue*, University of Washington; Ruth Etzioni, Fred Hutchinson Cancer Research Center; Elizabeth Slate, Medical University of South Carolina; Christopher Morrell, Loyola College in Maryland; David F. Penson, Keck School of Medicine and University of Southern California/Norris Cancer Center

9:45 Bayesian Networks for Prostate Cancer Modeling
Bradley M. Broom*, University of Texas M. D. Anderson Cancer Center; Devika Subramanian, Rice University

10:10 Floor Discussion

79. STATISTICAL METHODS IN QUANTITATIVE GENETICS AND GENOMICS *Salon G*

Sponsor: ENAR
Organizer: Jaya M. Satagopan, Memorial Sloan-Kettering Cancer Center
Chair: Jaya M. Satagopan, Memorial Sloan-Kettering Cancer Center

8:30 Bayes Estimators for Molecular Genealogy
Bruce Walsh*, University of Arizona
9:00 Lists of Lists: A Hierarchical Inference Problem from Genomics
Michael A. Newton*, University of Wisconsin–Madison
9:30 Class Discovery and Classification of Tumor Samples Using Mixture Modeling of Gene Expression Data—A Unified Approach
Shili Lin*, Ohio State University
10:00 Floor Discussion

80. STATISTICAL METHODS FOR ANALYSIS OF GENE-ENVIRONMENT INTERACTION *Salon K*

Sponsors: ASA Section on Epidemiology/ASA Section on Survey Research Methods
Organizer: Nilanjan Chatterjee, National Cancer Institute and Clarice Weinberg, National Institute of Environmental Health Sciences
Chair: Michael P. Epstein, Emory University

8:30 Semiparametric Methods for Estimating Gene-Environment Interaction Parameters from Case-Control Studies
Christie Spinka, University of Missouri–Columbia; Raymond Carroll, Texas A&M University; Nilanjan Chatterjee*, National Cancer Institute
9:00 A Method for Using Nuclear Families to Identify Gene by Environment Interaction
Emily O. Kistner* and Clarice R. Weinberg, National Institute of Environmental Health Sciences; Claire Infante-Rivard, McGill University
9:30 Haplotypes in Studies of Gene x Environment Interaction
Peter Kraft*, Harvard University
10:00 Discussant: Mitchell Gail, National Cancer Institute

81. ISOTONIC METHODS IN TOXICOLOGY & RISK *Salon B*

Sponsors: ASA Risk Analysis Section/ASA Biometrics Section
Organizer: Laura H. Gunn, Georgia Southern University
Chair: Laura H. Gunn, Georgia Southern University

8:30 Bayesian Methods for Assessing Ordering in Hazard Functions
Laura H. Gunn*, Georgia Southern University; David B. Dunson, National Institute of Environmental Health Sciences
9:00 Use of Historical Controls in Survival-Adjusted Quantal Response Tests for Comparing Tumor Incidence Rates
Shyamal D. Peddada*, Gregg E. Dinse, and Grace E. Kissling, National Institute of Environmental Health Sciences
9:30 Using Isotonic Regression to Identify the Ideal Recall Rate in Screening Mammography
Michael J. Schell*, Bahjat F. Qaqish, and Bonnie C. Yankaskas, University of North Carolina at Chapel Hill; Monique A. Amamoo, Shaw University; William E. Barlow, University of Washington
10:00 Discussant: Walter Piegorsch, University of South Carolina

82. NOVEL ENVIRONMENTAL APPLICATIONS OF SPATIAL STATISTICS *Salon D*

Sponsors: IMS/ASA Section on Statistics and the Environment
Organizer: Stephen L. Rathbun, Pennsylvania State University
Chair: Stephen L. Rathbun, Pennsylvania State University

8:30 Single- and Multi-Resolution Coregionalized Models for Spatially-Varying Growth Curves
Sudipto Banerjee* and Gregg A. Johnson, University of Minnesota
8:55 Spatial Analysis of Sea Turtle Nesting at Juno Beach, Florida
Lance A. Waller*, Traci Leong, and Andrew Barclay, Emory University; Bud Howard, Department of Environmental Resources Management, Palm Beach County
9:20 Modeling Multivariate Spatial Variables
Hao Zhang*, Washington State University
9:45 Hierarchical Bayesian Modeling of Invasive Species
Christopher K. Wikle* and Mevin B. Hooten, University of Missouri–Columbia
10:10 Floor Discussion

83. CONTRIBUTED PAPERS: NONPARAMETRICS
Meeting Room 415

Sponsor: ENAR
Chair: Randall H. Rieger, West Chester University

8:30 Comparison of Curves Based on a Cramer-von-Mises Statistic
Hua Liang*, St. Jude Children's Research Hospital
8:45 Canonical Correlates for Four Sets of Functional Data Curves
Peter M. Meyer* and Sue Leurgans, Rush University Medical Center
9:00 On Some Tests of the Covariance Matrix Under General Conditions
Arjun K. Gupta, Bowling Green State University; Jin Xu*, University of California, Riverside
9:15 Nonparametric Estimation of Stable Exponent
Zhaozhi Fan*, University of New Hampshire
9:30 Nonparametric Deconvolution, Traditional and Nontraditional Pooled Designs
Albert Vexler*, Aiyi Liu, and Enrique F. Schisterman, National Institute of Child Health and Human Development, National Institutes of Health
9:45 Pointwise Comparisons for Functional Data Inference
Dennis Cox and Jong Soo Lee*, Rice University
10:00 Simultaneous Estimation of Individual Patients' Responses
Jin Zhu*, Novartis Pharmaceuticals

84. CONTRIBUTED PAPERS: STATISTICAL METHODS IN SCREENING
Meeting Room 404

Sponsor: ENAR
Chair: Aniko Szabo, Huntsman Cancer Institute

8:30 Statistical Evaluation of Internal and External Mass Calibration Laws Utilized in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry
Ann L. Oberg*, David C. Muddiman, Terry M. Therneau, and Jeanette E. Eckel-Passow, Mayo Clinic and Foundation
8:45 On Criteria for Evaluating Models of Absolute Risk
Ruth M. Pfeiffer* and Mitchell H. Gail, National Cancer Institute, NIH
9:00 Assessing Relative Accuracy of Screening Tests in a Randomized Paired Screen Positive Design
Todd A. Alonzo*, University of Southern California; John M. Kittelson, University of Colorado Health Sciences Center
9:15 Modeling the Relationship Between Sensitivity and Sojourn Time in Periodic Cancer Screening
Dongfeng Wu*, Mississippi State University; Gary L. Rosner and Lyle D. Broemeling, University of Texas M. D. Anderson Cancer Center
9:30 Estimation of Sensitivity and Sojourn Time in Breast Cancer Screening Studies
Xiuyu J. Cong*, Rice University; Yu Shen, University of Texas M. D. Anderson Cancer Center
9:45 Floor Discussion

85. CONTRIBUTED PAPERS: MEASUREMENT ERROR
Salon E

Sponsor: ENAR
Chair: Chin-Shang Li, St. Jude Children's Research Hospital

8:30 A Linear Mixed Model with Heteroscedastic Covariate Measurement Error
Liang Li* and Tom Greene, Cleveland Clinic Foundation
8:45 **Latent Class Regression on Latent Factors**
Jia Guo* and Melanie M. Wall, University of Minnesota; Yasuo Amemiya, IBM T. J. Watson Research Center
9:00 **Model Robustness in Structural Measurement Error Modeling**
Xianzheng Huang*, Leonard A. Stefanski, and Marie Davidian, North Carolina State University
9:15 On Corrected Score Approach for Proportional Hazards Model with Covariate Measurement Error
Xiao Song*, University of Washington; Yijian Huang, Emory University
9:30 A Bayesian Adjustment of Covariate Misclassification with Correlated Binary Outcome Data
Dianxu Ren* and Roslyn A. Stone, University of Pittsburgh
9:45 Evaluating and Correcting Guess Effect in Imperfect Double-Blinded Clinical Trials
Jianfeng Cheng* and Eva Petkova, Columbia University
10:00 A Latent Variable Model for Measurement Error Correction Using Replicate Data
Sohee Park*, Louise M. Ryan, John Meeker, and Russ Hauser, Harvard University

86. CONTRIBUTED PAPERS: FRAILTY MODELS *Salon J*

Sponsor: ENAR
Chair: Hongmei Zhang, University of West Florida

8:30 A Shared Gamma Frailty Model for Dependent Failure and Truncation Times
Emily C. Martin* and Rebecca A. Betensky, Harvard University

8:45 Parametric Frailty Models for Quality of Life in Oncology
Andrea B. Troxel*, University of Pennsylvania School of Medicine

9:00 **Likelihood Ratio Test for the Variance Component in a Semi-Parametric Shared Gamma Frailty Model**
Xin Zhi*, Eli Lilly and Company; Lynn Eberly and Patricia Grambsch, University of Minnesota

9:15 Bayesian Semiparametric Dynamic Frailty Models for Multiple Event Time Data
Michael L. Pennell*, University of North Carolina at Chapel Hill and National Institute of Environmental Health Sciences; David B. Dunson, National Institute of Environmental Health Sciences

9:30 Extensions of Multivariate Survival Analysis to Include Grouped Failure Time Data with Application in Quality of Life
Denise A. Esserman*, Columbia University; Andrea B. Troxel, University of Pennsylvania School of Medicine

9:45 A Joint Frailty Model for Survival Time and Gap Times Between Recurrent Events
Xuelin Huang*, University of Texas M. D. Anderson Cancer Center; Lei Liu, University of Virginia School of Medicine

10:00 A Stratified Frailty Gap Time Model Fitted by a Re-Censoring Method
Lei Liu*, University of Virginia School of Medicine; Xuelin Huang, University of Texas M. D. Anderson Cancer Center

87. CONTRIBUTED PAPERS: MULTIPLE IMPUTATION *Salon F*

Sponsor: ENAR
Chair: Anthony Almudevar, University of Rochester

8:30 Multiple Imputation for Non-Normal Missing Data Using Tukey's gh Distribution
Yulei He* and Trivellore E. Raghunathan, University of Michigan

8:45 Multiple Imputation for Marginal and Linear Mixed Effects Models in the Analysis of Longitudinal Data with Informative Missingness
Wei Deng* and Lei Shen, Ohio State University

9:00 Small Sample and Asymptotic Relationships Between Multiple Imputation, Maximum Likelihood, and Fully Bayesian Methods
Qingxia Chen* and Joseph G. Ibrahim, University of North Carolina at Chapel Hill

9:15 Estimating the Dose Response Relationship for Occupational Radiation Exposure Measured with Minimum Detection Level
Xiaonan (Nan) Xue*, Albert Einstein College of Medicine of Yeshiva University; Roy E. Shore, New York University School of Medicine; Mimi Y. Kim, Albert Einstein College of Medicine of Yeshiva University; Xiangyang Ye, New York University School of Medicine

9:30 Floor Discussion

88. CONTRIBUTED PAPERS: GENE EXPRESSION ANALYSIS IN CANCER RESEARCH *Meeting Room 412*

Sponsor: ENAR
Chair: Ram C. Tiwari, National Cancer Institute

8:30 A Comparison of Gene Expression Measurements from Commercial Microarray Platforms
Karla V. Ballman*, Christopher P. Kolbert, and Sreekumar Raghavakaimal, Mayo Clinic College of Medicine

8:45 **A Two-Stage Mixture Model Strategy for Meta-Analysis of Microarray Data**
Ronglai Shen*, Debashis Ghosh, and Arul M. Chinnaiyan, University of Michigan

9:00 Bayesian Covariance Selection
Adrian Dobra*, Duke University

9:15 Model Selection Techniques in Gene Expression Profiling for Predicting Breast Cancer Outcome
Zhaoling Meng* and Bret Musser, Merck & Co., Inc.

9:30 Analysis of Gene Expression Data Using Split-Merge Markov Chain Monte Carlo
Sonia Jain*, University of California, San Diego; Radford M. Neal, University of Toronto

9:45 Constructing Prognostic Gene Signatures for Cancer Survival
Derick R. Peterson*, University of Rochester Medical Center
10:00 Floor Discussion

Wednesday, March 23
10:15–10:30 a.m.

Break *Grand Ballroon Pre-Function Area*

Wednesday, March 23
10:30 a.m.–12:15 p.m.

89. BAYESIAN AND NON-BAYESIAN APPROACHES TO COMPETING RISKS *Salon A*

Sponsors: ENAR/ASA Section on Statistics in Defense and National Security/ASA Risk Analysis Section
Organizer: Joseph G. Ibrahim, University of North Carolina at Chapel Hill
Chair: Joseph G. Ibrahim, University of North Carolina at Chapel Hill

10:30 A Marginal Conditional Model for Multivariate Survival Data
Glen A. Satten*, Centers for Disease Control and Prevention; Somnath Datta, University of Georgia
11:00 Bayesian Analysis of Competing Risks in Cancer Survival
Sanjib Basu*, Northern Illinois University and Rush University Medical Center
11:30 Joint Competing Risk Modeling for Assessing Important PSA Markers in Predicting Prostate Cancer Specific Mortality
Ming-Hui Chen*, University of Connecticut; Joseph G. Ibrahim, University of North Carolina at Chapel Hill; Anthony V. D'Amico, Harvard University
12:00 Floor Discussion

90. INTEGRATING MULTIPLE SOURCES OF GENOMIC DATA *Salon G*

Sponsor: ENAR
Organizer: Mahlet G. Tadesse, University of Pennsylvania
Chair: Mahlet G. Tadesse, University of Pennsylvania

10:30 Statistical Methods for ChIP-chip High-Density Oligonucleotide Array Data
Sunduz Keles*, University of Wisconsin–Madison; Mark J. van der Laan and Sandrine Dudoit, University of California, Berkeley; Simon E. Cawley, Affymetrix
11:00 Improving the Accuracy of Protein-Protein Interaction Network Using Local Graph Structure and Comparative Genomics
Peter J. Park* and Jung-Ah Lim, Children's Hospital, Boston
11:30 Pathway-Based Analysis of DNA Microarray Data
Marina Vannucci*, Texas A&M University
12:00 Floor Discussion

91. BIOSURVEILLANCE GEOINFORMATICS FOR BIOSECURITY (THEME SESSION) *Salon B*

Sponsors: ENAR/ASA Section on Statistics in Defense and National Security/ASA Section on Survey Research Methods
Organizer: G. P. Patil, Pennsylvania State University
Chair: G. P. Patil, Pennsylvania State University

10:30 Approaches for Reducing Spurious Cluster Identification in Scan Statistics
Howard S. Burkom*, Johns Hopkins University
11:00 Bayesian Spatial Surveillance of Small Area Disease Events: Particle Filtering Methods
Carmen L. Vidal Rodeiro and Andrew B. Lawson*, University of South Carolina
11:30 Biosurveillance Geoinformatics of Hotspot Detection and Prioritization for Biosecurity
G.P. Patil and Stephen L. Rathbun, Pennsylvania State University
12:00 Floor Discussion

92. RISK RANKING AND DISEASE MAPPING *Salon D*

Sponsors: ASA Risk Analysis Section/ASA Health Policy Statistics Section
Organizers: Xiao-Li Meng, Harvard University, and Vanja Dukic, University of Chicago
Chair: Xiao-Li Meng, Harvard University

10:30 Type S and Type M Error Rates for Ranking Comparisons
Samantha R. Cook* and Andrew Gelman, Columbia University; Francis Tuerlinckx, University of Leuven
11:00 Optimal Survival Curve Ranking (OSCR): Application to Assessment of AIDS Reporting Delay
Vanja Dukic*, University of Chicago; Peter Bouman, Northwestern University; Xiao-Li Meng, Harvard University
11:30 Bayesian Ranking Methods with Applications to Disease Mapping
Thomas A. Louis* and Rongheng Lin, Johns Hopkins University; Susan M. Paddock and Greg Ridgeway, Rand Corporation
12:00 Floor Discussion

93. TO MIX OR NOT TO MIX *Salon E*

Sponsors: ENAR/ASA Section on Statistics in Defense and National Security
Organizer: Ramani S. Pilla, Case Western Reserve University
Chair: Richard Charnigo, University of Kentucky

10:30 Interpretation of a Mixture Model
Bruce G. Lindsay*, Pennsylvania State University; Surajit Ray, Statistical and Applied Mathematical Sciences Institute (SAMSI)
11:00 Remarks on Mixtures of Regressions
David W. Scott*, Rice University
11:30 Local Likelihoods versus Local Mixture Likelihoods
Ramani S. Pilla* and Catherine Loader, Case Western Reserve University
12:00 Floor Discussion

94. CONTRIBUTED PAPERS: STATE SPACE MODELS AND TIME SERIES ANALYSIS *Meeting Room 415*

Sponsor: ENAR
Chair: Lei Shen, Ohio State University.

10:30 State Space Models of Immune Responses Under Treatment in Plasma and Lymph Nodes
Wai-Yuan Tan, University of Memphis; Ping Zhang*, Middle Tennessee State University; Xiaoping Xiong, St. Jude Research Hospital
10:45 Random Coefficient Transfer Function Model for Panel Data
Hyunyoung Choi* and Hernando Ombao, University of Illinois at Urbana–Champaign; Bonnie Ray, IBM T. J. Watson Research Center
11:00 Investigation of Synchrony Between Brain Regions Using Wavelet Coherence
Bing Gao*, Hernando Ombao, and Christopher Edgar, University of Illinois at Urbana-Champaign
11:15 Automated Peak Identification in a Time-of-Flight Spectrum
Haijian Chen*, Eugene R. Tracy, William E. Cooke, and Michael W. Trosset, College of William and Mary
11:30 Variance-Covariance Estimation with an Application to Intercellular Signaling
Scott H. Holan*, University of Missouri–Columbia
11:45 Floor Discussion

95. CONTRIBUTED PAPERS: CAUSAL INFERENCE *Meeting Room 412*

Sponsor: ENAR
Chair: Jennie Ma, University of Texas Health Science Center at San Antonio

10:30 Neuropathologic Mediators of the Association Between the Apolipoprotein E epsilon4 Allele and Clinical Dementia
Yan Li*, Julia L. Bienias, and David A. Bennett, Rush University Medical Center
10:45 Artificial Censoring for Randomized Clinical Trials in the Presence of Non-Random Non-Compliance
Long-Long Gao* and Marshall M. Joffe, University of Pennsylvania School of Medicine
11:00 **Causal Inference in Hybrid Intervention Trials Involving Treatment Choice**
Qi Long*, Roderick J. A. Little, and Xihong Lin, University of Michigan
11:15 Polydesigns in Causal Inference
Fan Li* and Constantine E. Frangakis, Johns Hopkins University

11:30 A Formal Approach for Defining and Identifying the Fundamental Effects of Exposures on Disease from Sets of Experiments Conducted on Populations of Non-Identical Subjects
Steven D. Mark*, National Cancer Institute

11:45 **Bounds on Causal Effects in Three-Arm Trials with Noncompliance**
Jing Cheng*, Dylan S. Small, University of Pennsylvania

12:00 A Distributional Approach for Causal Inference Using the Propensity Score
Zhiqiang Tan*, Johns Hopkins University

96. CONTRIBUTED PAPERS: CLINICAL TRIALS II

Salon K

Sponsor: ENAR
Chair: Vladimir Dragalin, GlaxoSmithKline

10:30 **Multi-center Clinical Trials: Randomization and Ancillary Statistics**
Lu Zheng* and Marvin Zelen, Harvard School of Public Health

10:45 Designs for Phase I Clinical Trials with Continuous/ Multinomial Toxicities
Zhilong Yuan*, Rick Chappell, University of Wisconsin–Madison

11:00 Sample Size Calculation in Survival Trials Accounting for Time-Dependent Dynamics of Noncompliance and Risk
Bingbing Li* and Patricia Grambsch, School of Public Health, University of Minnesota

11:15 Multi-Center Trials with Binary Response
Vladimir Dragalin and Valerii Fedorov*, GlaxoSmithKline

11:30 How to Prepare the Best Data Package for a Data Monitoring Committee
Vipin Arora*, Novartis Pharmaceutucal Corporation; David Manner, Eli Lilly and Company

11:45 Evaluation of the Quality of Investigative Centers Using Clinical Rating and Compliance Data
Junyuan Wang* , Merck Research Laboratory; Junfeng Sun, Ohio State University; Guanghan Liu, Merck Research Laboratory

12:00 Detecting Qualitative Interactions in Clinical Trials: An Extension of Range Test
Jianjun (David) Li* and Ivan Chan, Merck

97. CONTRIBUTED PAPERS: SPATIAL MODELING

Meeting Room 404

Sponsor: ENAR
Chair: John Castelloe, SAS Institute

10:30 Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets
Cari Kaufman* and Mark Schervish, Carnegie Mellon University; Doug Nychka and Reinhard Furrer, National Center for Atmospheric Research

10:45 Recursive Partitioning in Spatially Correlated Data
Patrick S. Carmack*, University of Texas Southwestern Medical Center; William R. Schucany, Southern Methodist University

11:00 Spatial Stochastic Volatility
Huiliang Xie* and Jun Yan, University of Iowa

11:15 **Bayesian Areal Wombling for Geographical Boundary Analysis**
Haolan Lu* and Brad Carlin, University of Minnesota

11:30 Simultaneous Confidence Intervals for Ratios of Nonparametric Intensity
Traci l. Leong* and Lance A. Waller, Rollins School of Public Health, Emory University

11:45 Process Convolution Approach to Reconstruction of Binary Fields
Margaret B. Short* and Dave Higdon, Los Alamos National Lab

12:00 Spatial Estimation with Computer-Efficient Parsimonious Interaction Models
Ernst Linder*, University of New Hampshire

98. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS III
Salon J

Sponsor: ENAR
Chair: Bin Wang, University of South Alabama

10:30 Application and Assessment of Residual-Based Classification Approaches on Melanoma Survival Data: Validation of AJC on Cancer Melanoma Staging System
Chen-An Tsai*, Dung-Tsa Chen, and Seng-Jaw Soong, UAB Comprehensive Cancer Center

10:45 Local Linear Estimation of a Smooth Distribution Based on Censored Data
Liang Peng, Georgia Institute of Technology; Shan Sun*, Texas Tech University

11:00 Multiple Augmentation with Outcome Dependent Sampling
Shuangge Ma*, University of Washington

11:15 Survival Model and Estimation for Lung Cancer Patients
Xingchen A. Yuan*, East Tennessee State University; Don Hong and Shyr Yu, Vanderbilt University

11:30 Modeling Religion's Influence on HIV Progression and Mortality
John A. Myers*, Musie Ghbremicheal, and Heping Zhang, Yale School of Medicine

11:45 On a Connection Between the Parametric Likelihood and the Empirical Likelihood
Min Chen* and Mai Zhou, University of Kentucky

12:00 One- and Two-Sample Nonparametric Inference Procedures in the Presence of Dependent Censoring
Yuhyun Park* , Harvard University; Lu Tian, Northwestern University; L. J. Wei, Harvard University

99. CONTRIBUTED PAPERS: ASSOCIATIVE ANALYSIS OF GENETIC DATA
Salon F

Sponsor: ENAR
Chair: Susan Halabi, Duke University

10:30 Incorporating Clustering Uncertainty in Regression-Based Analysis for Haplotype-Disease Association
Jung-Ying Tzeng*, North Carolina State University

10:45 Using Tree-Based Recursive Partitioning Methods to Group Haplotypes in Association Studies
Kai Yu*, Washington University, St. Louis; Jun Xu, Procter & Gamble Co.; D. C. Rao and Michael Province, Washington University, St. Louis

11:00 Inference Under Various Multinomial Distribution Models for Haplotypes Under Hardy-Weinberg Disequilibrium
Beverly M. Snively* and David M. Reboussin, Wake Forest University School of Medicine; Christine E. McLaren, University of California, Irvine; Ronald T. Acton, University of Alabama at Birmingham; Mark R. Speechley, University of Western Ontario; Emily L. Harris, Kaiser Permanente Northwest; James C. Barton, Southern Iron Disorders Center; Cathie Leiendecker-Foster, University of Minnesota; Victor R. Gordeuk, Howard University

11:15 Inference on Haplotype-Environment Interactions Using Genotype Data from Case-Control Studies
Lydia C. Kwee*, Amita K. Manatunga, Emory University ;Glen A. Satten, Centers for Disease Control and Prevention; Michael P. Epstein, Emory University

11:30 Joint Linkage and Association Mapping of Quantitative Trait Loci, A Haplotype-Based Approach
Ruzong Fan*, Texas A&M University; Jeesun Jung, University of Pittsburgh; Lei Jin, Texas A&M University

11:45 Quantifying Bias Due to Genotyping Error in Case-Control Studies of Gene Haplotypes and Cancer
Usha S. Govindarajulu*, Donna Spiegelman, David J. Hunter, and Peter Kraft, Harvard University

12:00 Accounting for Population Stratification in Case-Control Studies of Genetic Association: A Bayesian Approach
Li Zhang*, Bhramar Mukherjee, Malay Ghosh, and Rongling Wu, University of Florida

ABSTRACTS

1. RECENT DEVELOPMENTS IN SEQUENTIAL CLINICAL TRIALS METHODOLOGY

DOSE FINDING BASED ON EFFICACY AND TOXICITY IN PHASE I/II CLINICAL TRIALS

Peter F. Thall*, University of Texas M. D. Anderson Cancer Center
John D. Cook, University of Texas M. D. Anderson Cancer Center

A new method for dose-finding in phase I/II clinical trials based on both efficacy and toxicity is presented. The method is outcome-adaptive, with the doses of successive cohorts of patients chosen based on the data from patients treated previously in the trial. The goal is to find a dose of an experimental agent that has both acceptably low toxicity and acceptably high efficacy. As a basis for comparing doses, the method uses an explicit trade-off between the probabilities of toxicity and efficacy, elicited graphically from the physician planning the trial. The method will be illustrated by two applications: a clinical trial of graft-versus-host disease prophylaxis in allogeneic bone marrow transplantation, and a trial of agents for rapid treatment of acute ischemic stroke. Computer simulations of the method in the context of these trials will be presented. (Biometrics 60:684-693, 2004)

email: rex@mdanderson.org

DESIGN OF GROUP SEQUENTIAL CLINICAL TRIALS WITH ORDINAL CATEGORICAL DATA

Jonathan J. Shuster, University of Florida
Lili Tian*, University of Florida
Myron Chang, University of Florida

This talk is dedicated and restricted to the design of group sequential clinical trials with ordinal categorical outcome data. In practice, most biostatisticians would seek to utilize the Mann-Whitney-Wilcoxon test as their test statistic. Although the design of such trials under a model, that the odds ratios for the cumulative distributions are identical, can be obtained from the combined work of Jones and Whitehead (1979) and Whitehead (1993), group sequential designs with arbitrary alternative hypotheses (i.e. those not satisfying the common odds ratio assumption for the cumulative distributions.) have not been well addressed. This talk will focus on this subject. This talk is based on a paper on theory (Shuster, Chang, Tian, sequential analysis, 2004) and a working paper on small sample results (Tian, Chang, Shuster).

email: ltian@biostat.ufl.edu

EFFICIENT GROUP SEQUENTIAL DESIGNS WHEN THERE ARE SEVERAL EFFECT SIZES UNDER CONSIDERATION

Chris Jennison, University of Bath, U.K.
Bruce W. Turnbull*, Cornell University

We consider the construction of efficient group sequential designs where the goal is a low expected sample size not only at the null hypothesis and the alternative (taken to be the minimal clinically meaningful effect size), but also at more optimistic anticipated effect sizes. Pre- specified Type I error rate and power requirements can be achieved both by standard group sequential tests and by more recently proposed adaptive procedures. We investigate four nested classes of designs: (A) Group sequential tests with equal group sizes and stopping boundaries determined by a monomial error spending function (the ``rho-family''); (B) As A but the initial group size is allowed to be different from the others; (C) Group sequential tests with arbitrary group sizes and arbitrary boundaries, fixed in advance; (D) Adaptive tests --- as C but at each analysis, future group sizes and critical values are updated depending on the current value of the test statistic. By examining the performance of optimal procedures within each class, we conclude that class~B provides simple and efficient designs with efficiency close to that of the more complex designs of classes C and D. We provide tables and figures illustrating the performances of optimal designs within each class and defining the optimal procedures of classes A and B.

email: bwt2@cornell.edu

---

2. CUTTING-EDGE RISK ASSESSMENT ISSUES AND METHODS

DEVELOPMENTAL NEUROTOXICITY MODELING AND RISK ASSESSMENT

Mehdi Razzaghi*, Bloomsburg University
Ralph L. Kodell, NCTR/FDA

The goal of developmental neurotoxicity study is to assess neurological development in laboratory animals following maternal exposure to a toxic substance. These studies are conducted to identify and characterize potential adverse effects of specific chemicals on the developing nervous system and are generally utilized to assess health risks in humans. The process of risk assessment in such studies requires careful consideration of several issues from design of the experiment to dose-response modeling and extrapolation to low exposure levels. Here, we consider statistical problems that occur in the design and analysis of developmental neurotoxicity experiments. A dose- response model based on a two-stage hierarchical normal structure is proposed and maximum likelihood estimation of the model parameters is discussed. Upper confidence limits on the excess risk for fixed low exposure levels as well as benchmark doses for fixed levels of excess risk are derived. An experimental data set that examines developmental toxicity of a mixture of two herbicides in rats is used for illustration of the proposed methodology.

email: razzaghi@bloomu.edu

APPLICATION OF RISK ASSESSMENT AND MODELING APPROACHES FOR EVALUATING GENE THERAPY RISKS

Steven Anderson*, CBER/FDA

Gene therapy and stem cell therapy hold potential promise for the treatment of certain diseases. Although there have been some successes there are potential risks associated with this emerging technology. Recently, two children in a gene therapy clinical trial for X-linked Severe Combined Immune Disease (X-SCID) were diagnosed with leukemia out of a total of 9 children treated. Further examination of cells from each patient indicated that the gene therapy vector had inserted in the genome near an oncogene, whose dysregulation may have given rise to the leukemia. Although the gene therapy treatment process is complex, we will show approaches for statistical and mathematical modeling of the therapeutic and biological processes involved. Among the methods used in the model are Monte Carlo analysis, parameterization of data, and uncertainty and sensitivity analysis. The goal of this work is to offer predictions of the probability of severe adverse events such as cancer. Furthermore, approaches for using the model to evaluate the effectiveness of risk reduction interventions such as decreasing the quantity of vector or stem cell dose, etc. and their effect on predicted adverse events will be discussed.

email: andersonst@cber.fda.gov

COMPARING MODEL AVERAGING WITH OTHER MODEL SELECTION STRATEGIES FOR RISK ESTIMATION

Matthew W. Wheeler*, National Institute for Occupational Safety and Health
A. John Bailer, Miami University

Bayesian model averaging (BMA) has been proposed as a method of accommodating model uncertainty when estimating risk. Although the use of this model averaging technique is inherently appealing, little is known about its performance under general modeling conditions. We investigate the use of BMA for estimating excess risk under a Monte Carlo simulation. Dichotomous response data are simulated under various assumed underlying dose-response curves, and nine dose-response models (from the USEPA BMD model suite) are fit to obtain both model specific and BMA risk estimates. The risk estimates from the BMA method, as well as estimates from other commonly selected models, e.g., best fitting model or the model resulting in the most conservative fit, are compared to the true risk value to better understand both bias and coverage behavior in the estimation procedure. The BMA typically provides an unbiased estimate of true risk having similar variability when compared to estimates derived from the assumed model. Further, when appropriate models are used to calculate the BMA, the lower bound estimate provided coverage at the nominal level, which is superior to the other strategies considered. This approach provides an alternative method for risk managers to estimate risk while incorporating model uncertainty.

email: aez0@cdc.gov

MODEL AVERAGING USING THE KULLBACK INFORMATION CRITERION IN ESTIMATING EFFECTIVE DOSES FOR MICROBIAL INFECTION AND ILLNESS

Hojin Moon*, NCTR/FDA
Ralph L. Kodell, NCTR/FDA
James J. Chen, NCTR/FDA
Hyun-Joo Kim, Truman State University
David W. Gaylor, Gaylor and Associates, LLC

Since the National Food Safety Initiative of 1997, risk assessment has been an important issue in food safety areas. Microbial risk assessment is a systematic process for describing and quantifying a potential to cause adverse health effects associated with exposure to microorganisms. Various dose-response models for estimating microbial risks have been investigated. We have considered four two-parameter models and four three- parameter models in order to evaluate variability among the models for microbial risk assessment using infectivity and illness data from studies with human volunteers exposed to a variety of microbial pathogens. Model variability is measured in terms of estimated ED01's and ED10's, with the view that these effective dose levels correspond to the lower and upper limits of the 1% to 10% risk range generally recommended for establishing benchmark doses in risk assessment. Parameters of the statistical models are estimated using the maximum likelihood method. In this paper a weighted average of effective-dose estimates from eight two- and three-parameter dose-response models, with weights determined by the Kullback information criterion, is proposed to address model uncertainties in microbial risk assessment. The proposed procedures for incorporating model uncertainties and making inferences are illustrated with human infection/illness dose-response data sets.

email: hmoon@nctr.fda.gov

3. ANALYSIS AND ISSUES IN MATCHED FAMILY AGGREGATION STUDIES

ADVANCES IN THE DESIGN AND ANALYSIS OF FAMILIAL AGGREGATION STUDIES

Abigail G. Matthews, Harvard University
Dianne M. Finkelstein, Massachusetts General Hospital
Rebecca A. Betensky*, Harvard University

Analysis of family studies of disease must take into account dependencies among family members due to shared genes and environment. Other salient features of family studies that must also be acknowledged in both the design and analysis are complex non-random ascertainment, varying family sizes, and small sample sizes. We describe recent advances on all of these fronts and apply our methodology to family studies of cancer.

email: betensky@hsph.harvard.edu

FAMILY-SPECIFIC APPROACHES TO THE ANALYSIS OF CASE-CONTROL FAMILY DATA

John Neuhaus*, University of California, San Francisco
Alastair Scott, University of Auckland, New Zealand
Chris Wild, University of Auckland, New Zealand

Case-control studies augmented by responses and covariates from family members allow investigators to more efficiently estimate the associations of interest in the original case-control sample as well as to directly relate within-family differences in covariates to differences in the response. Existing approaches for case-control family data parametrize covariate effects in terms of the marginal probability of response and measure the same covariate effects that one estimates from standard case-control studies. However, estimates of within-family covariate effects are often of greater scientific interest. This talk presents a profile likelihood approach that applies generally to settings where one has a fully specified model for the vector of responses in a family and particularly to family-specific models such as binary mixed-effects models. We will illustrate the approach using data from a case-control family study of brain cancer and consider the role of conditional likelihood methods.

email: john@biostat.ucsf.edu

A SEMIPARAMETRIC METHOD FOR ANALYZING MATCHED CASE-CONTROL FAMILY STUDIES

Molin Wang*, Harvard University and Dana-Farber Cancer Institute
John M. Williamson, National Center for Infectious Diseases, Centers for Disease Control and Prevention
Susan Redline, Rainbow Babies and Children's Hospital and Case Western Reserve University

We consider matched case-control familial studies which match a group of patients, called `case probands', with a group of disease-free subjects,called `control probands', using a set of family-level matching variables. Family members of each proband are then recruited into the study. Of interest here is the familial aggregation of the response variable and the effects of subject-specific covariates on the response. We propose an estimating equation approach to jointly estimate the main effects and intrafamilial correlations for matched family studies with a continuous outcome. Only knowledge of the first two joint moments of the response variable is required. The induced estimators for the main effects and intrafamilial correlations are consistent and asymptotically normally distributed. We apply the proposed method to sleep apnea data. A simulation study demonstrates the usefulness of our approach.

email: mwang@jimmy.harvard.edu

4. BIOMETRICS SPECIAL INVITED PAPER SESSION

RECENT DEVELOPMENTS IN COMPUTATIONAL BIOLOGY

Wing Hung Wong* and Hongkai Ji, Stanford University

We outline several questions of central interest to current biology and review the statistical and computational challenges arising from the use of genome sequences and microarray data to study these questions. First we review methods for analyzing genome sequences with emphasis on identifying and working with conserved regions. Then we review classification and clustering methods in the analysis of microarray data, and discuss how these analyses can help to extract candidate sets of co-regulated genes, and how additional microarray-based assays such as ChIP-chip analysis may provide information on the mechanism of regulation. Finally, we discuss approaches that combine all the above sources of information to elucidate transcriptional regulatory networks.

email: whwong@stanford.edu

5. LACK OF FIT TESTS FOR MODEL MISSPECIFICATION WITH APPLICATIONS TO LONGITUDINAL DATA AND SURVIVAL DATA ANALYSIS

USING BAYESIAN STATISTICS TO TEST FOR LACK OF FIT IN FREQUENTIST FASHION

Jeffrey D. Hart*, Texas A&M University

The Bayesian paradigm is used to construct test statistics of the null hypothesis that a function lies in a parametric family. The statistics are nonparametric in the sense that the dimension of the alternative family grows without bound as the number of observations tends to infinity. Each statistic is simply the posterior probability of the null hypothesis corresponding to specified prior probabilities.

The frequentist may use such a posterior probability as a test statistic by determining its distribution under the null hypothesis and then rejecting the null when said probability is sufficiently small. BIC approximations to the null posterior probability are particularly attractive to the frequentist since they do not depend upon prior probabilities. We provide results on the asymptotic null distribution of a BIC probability, and compare the power of the corresponding test with that of other nonparametric lack-of-fit tests.

email: hart@stat.tamu.edu

A GOODNESS OF FIT TEST FOR PROPORTIONAL ODDS MODELS FOR SURVIVAL DATA

Linxu Liu*, Columbia University
Jianhua Huang, University of Pennsylvania

The proportional odds model provides a useful alternative to the popular Cox proportional hazards model in survival analysis. It assumes that the odds ratio of two individuals is constant over time provided that their covariate values do not change. However, this assumption needs to be justified in real data analysis. So far, formal statistical testing of this assumption is only limited to the two sample case. We propose a class of time-varying regression coefficient linear cumulative-odds models that includes the proportional odds model as a special case. Within this modeling class, the proportional odds hypothesis can be tested and when it is rejected, a plausible alternative model is provided. Estimation and inference procedures for the proposed model class are discussed. We illustrate our methods by applying it to node-positive primary breast cancer data.

email: ll2255@columbia.edu

THE IOS TEST FOR MODEL MISSPECIFICATION

Brett Presnell*, University of Florida
Marinela Capanu, University of Florida
Dennis D. Boos, North Carolina State University

The in-and-out-of-sample (IOS) test of model misspecification is based on the ratio of in-sample and out-of-sample likelihoods. The test is broadly applicable, and in simple problems it often approximates well known, intuitive test procedures. The IOS test statistic is asymptotically equivalent to a multiplicative contrast between two estimates of the model information matrix, a fact which leads to a proof of asymptotic normality while also revealing a familial relationship to the information matrix (IM) test popular in econometrics. Asymptotically valid p-values for the IOS test can be computed using the parametric bootstrap, and these bootstrap p-values are generally more accurate than those based directly on the normal approximation. The resulting methodology is demonstrated with a variety of examples and simulations involving both discrete and continuous data.

email: presnell@stat.ufl.edu

GOODNESS-OF-FIT TEST FOR MODEL ASSUMPTIONS IN LONGITUDINAL DATA

Annie Qu*, Oregon State University

Model selection and goodness-of-fit tests for model assumptions are important and challenging in longitudinal data analysis. Since the likelihood is often unknown or difficult to specify for complex and high dimensional data, traditional likelihood ratio tests are not obvious or easily applied in nonparametric approaches, estimating equation approaches or non-normal random effects models. Non-nested hypothesis testing might also occur in these situations. We will illustrate how to perform this type of test without the likelihood function.

email: qu@stat.orst.edu

6. COMPUTATIONAL METHODS

PARTIAL PRIOR IMPROVING NPML ESTIMATION FOR MIXTURES

Ji-Ping Z. Wang*, Northwestern University

Given observations originating from a mixture distribution $f[x;Q(\lambda)]$ where the kernel f is known and the mixing distribution Q is unknown, we consider to estimate a functional of Q. A natural estimator of such a functional, say $\theta(Q)$, can be obtained by substituting Q with its nonparametric maximum likelihood estimator (NPMLE), denoted here as $\hat Q$. We demonstrate however, that the plug-in estimator $\theta(\hat Q)$ can be unstable or substantially biased due to large variability of $\hat Q$ or structural properties of the parameter space of λ. In this paper we propose an approach using a \emph{partial prior} for Q in the form of a prior distribution for certain functional(s) of Q, e.g., $p[\theta(Q)]$ for variance and bias reduction purposes in a class of wide-ranging models. An empirical Bayes idea for choice of hyper-parameters in the prior is proposed. Optimization result and its realization in a VDM/ECM algorithm are discussed. The effectiveness of the adaptive penalizing approach in bias or variance reduction is illustrated by motivating examples of Binomial mixtures.

email: jzwang@northwestern.edu

WAVELETS AND EVOLUTION ALGORITHMS FOR MASS SPECTROMETRY DATA PROCESSING

Ming Li*, Vanderbilt University
Huiming Li, Vanderbilt University
Johnathan Xu, Vanderbilt University
Yu Shyr, Vanderbilt University
Don Hong, East Tennessee State University and Vanderbilt University

Mass spectrometric (MS) data holds invaluable information leading to disease diagnosis and treatment. The processing goal is to effectively and correctly obtain the true information from the raw MS data for further statistical analysis. Two general approaches have been studied recently: functional data analysis approach (Morris and Carroll 2004, Billheimer 2004) and the feature extraction approach (Coombes et al 2004, Chen, Hong and Shyr 2004). To provide a final peak list for future statistical analysis, the whole processing procedure by feature extraction approach usually takes the following steps: de- noising (smoothing), baseline correction, normalization, peak detection and alignment. In this paper, we deal with a real spectrometric data file by going through three processing procedures respectively: (1) GG method, Gaussian smoothing and Genetic algorithm for final peak binning, (2) WW method, Wavelet denoising and Window-based peak alignment, and (3) WG method, Wavelet denoising and Genetic algorithm for peak binning. The results show that wavelets are powerful in denoising and the genetic algorithm performs well in peak extraction. Therefore, we propose a framework based on the WG method for MS data processing.

email: ming.li@vanderbilt.edu

RECENT PROGRESS ON MASS SPECTROMETRY DATA PROCESSING

Don Hong*, East Tennessee State University and Vanderbilt University
Yu Shyr, Vanderbilt University

Mass spectrometry (MS) becomes one of the critical components in cancer research recently. However, there are many challenges both in MS data processing and data analysis. In this talk, we present some recent progress on MS data processing using mathematical tools and statistical methods. The tackled problems in the data processing procedure include standardizing data read-in, denoising, baseline correction, normalization, and peak determination. Some experimental results will be shown using the data processing software packages developed by Biostatistics/ Bioinformatics MS data research group in Vanderbilt Ingram Cancer Center.

email: don.hong@vanderbilt.edu

STOCHASTIC SIMULATION OF E. COLI 0157:H7 INFECTION IN CATTLE

Baktiar Hasan*, University of Guelph
Brian Allen, University of Guelph
Scott A. McEwen, University of Guelph

Escherichia coli O157:H7 is an important pathogen of humans that is usually contracted through contaminated food, but waterborne transmission and person-to-person spread may also occur. Infection also occurs in other species of animals, although usually not in association with disease. In North America, cattle are believed to be the principal reservoir of this organism for humans. It lives mainly in the gastrointestinal tract but can survive outside the bovine or human host for variable periods of time. Meat may become contaminated from carrier animals at the time of slaughter. Options for controlling this infection at the farm level include vaccination, use of feed additives, and segregation of animals during production or transport to slaughter. The dynamics of infection in cattle are poorly understood, and cattle production is complex. These and other factors limit the usefulness of traditional experimental approaches to assessing the potential impact of interventions. The objective of this research was to develop a simulation model of E. coli O157:H7 infection in cattle for use as a research tool.

email: ballen@uoguelph.ca

MOMENT ESTIMATORS FOR MUTATION RATES

Loki Natarajan*, University of California, San Diego
Charles C. Berry, University of California, San Diego
Christoph Gasche, University of California, San Diego

Spontaneous or randomly occurring mutations play a key role in cancer progression. Estimating the mutation rate of cancer cells can provide useful information about the disease. In this investigation, we develop a discrete-time stochastic model for a mutational birth process. We assume that mutations occur concurrently with mitosis, so that when a non-mutant parent cell splits into two progeny, one of these daughter cells could carry a mutation. A moment estimator is proposed for the mutation rate. Statistical properties (bias and mean-squared error) of this estimator are investigated via theory and simulations. Sensitivity of the proposed estimator to deviations from modeling assumptions are also explored. A salient feature of this estimator is the ease with which it can be computed. The methods developed will be applied to data from a human colorectal cancer cell line.

email: loki@math.ucsd.edu

A UNIFIED APPROACH FOR SIMULTANEOUS CLUSTERING AND DIFFERENTIAL EXPRESSION IDENTIFICATION

Ming Yuan*, Georgia Institute of Technology;
Christina Kendziorski, University of Wisconsin–Madison

Although both clustering and differentially expressed gene identification are equally essential in most microarray studies, the two tasks are often conducted without regard to each other. This is clearly not the most efficient way of extracting information. It is our main aim of this work to develop a coherent statistical method which can simultaneously cluster and detect differentially expressed genes.

email: yuanm@stat.wisc.edu

MODELING P-VALUES IN HIGH DIMENSIONAL TESTING APPLICATIONS USING A UNIFORM - BETA MIXTURE: THE PERFORMANCE OF INTERVAL ESTIMATES

Qinfang Xiang, University of Missouri–Rolla
Gary L. Gadbury*, University of Missouri–Rolla
Jode Edwards, USDA ARS and Iowa State University

Highly dimensional biological experiments commonly aim to identify variables that differ among groups of experimental units. In such experiments, statistical tests may be simultaneously conducted for each of several thousand hypotheses. Recently some investigators analyzing data from microarray experiments have found that the distribution of p-values from tests for differential genetic expression can contain useful information regarding several quantities of interest. A mixture of a beta distribution and a uniform distribution has been used to model the distribution of p- values. The resulting fitted model can be used to compute derived quantities such as a true positive (TP) probability. This presentation focuses on the precision of estimates from such a mixture model and compares computational methods for computing confidence intervals for model parameters. The role of the number of tests (i.e., number of p-values) in the precision of estimates is also considered. Markov – Chain Monte Carlo methods seemed to have more advantages over two other methods considered. Interestingly, the number of p-values must be larger than anticipated to produce precise interval estimates.

email: gadburyg@umr.edu

NONPARAMETRIC TESTS FOR DEPENDENT OBSERVATIONS OBTAINED AT VARYING TIME POINTS

Susanne May*, University of California, San Diego
Victor DeGruttola, Harvard University

We propose two methods for two-group comparisons of repeated measures of a response where the repeated measures might be obtained at arbitrary time points that differ over individuals. The tests are almost U-Statistics in that the kernel contains some unknown parameters that are estimated from the data. Our methods are appropriate for alternatives in which response means of one group are strictly greater than the response means of the other group. The first method does not make any assumption regarding the distributions of the repeated measures except that they not depend on the response values. The second method assumes that the repeated measures can be grouped into distinct periods of observations (e.g. around fixed follow-up time points) such that the covariance between scores only depends on the periods the observations belong to. Inference can conveniently be based on resampling. For certain alternative hypotheses, asymptotic normality of the test statistics can be shown. To investigate the properties of these methods, we perform a simulation study. We apply both methods to assess differences in HIV-1 RNA decline for drug resistant and drug sensitive patients.

email: smay@ucsd.edu

NONPARAMETRIC REGRESSION SUBJECT TO A MONOTONICITY CONSTRAINT

Matthew J. Schipper*, University of Michigan
Jeremy Taylor, University of Michigan
Xihong Lin, University of Michigan

Normal tissue complications are a possible side effect of radiation therapy and are related to the dose of radiation received by the normal tissue. The dose is often measured spatially in a tissue. It is not known what function of the dose distribution to the normal tissue drives the presence and severity of the complications. One summary measure is obtained by integrating a weighting function of dose (w(d)) over the dose distribution. A linear weight function (w(d)=d) corresponds to the commonly used mean dose model. For biological reasons the weight function should be smooth and monotonic. We propose to study the dose effect on a clinical outcome using a nonparametric method by estimating this weight function smoothly and subject to the monotonicity constraint. In our model w(d) is written as a cumulative sum of a smooth positive function of d. We illustrate our method with data from a head and neck cancer study in which the normal tissue complication is loss of saliva flow following radiation delivered to the parotid gland. The analysis is complicated by non-Gaussian hierarchical longitudinal data. We fit the model using ML and MCMC methods.

email: mjschipp@umich.edu

ON MINIMAX WAVELET ESTIMATOR WITH CENSORED DATA

Linyuan Li*, University of New Hampshire

Wavelet-based density estimators with randomly right- censored data are considered. We investigate the asymptotic rates of convergence of estimators based on thresholding of empirical wavelet coefficients. Our technique is facilitated by a result of Stute (1995) that approximates the Kaplan-Meier integrals as an average of i.i.d. random variables with a certain rate. We show that a block thresholded wavelet estimator achieves exactly optimal minimax convergence rates over a large range of Besov function classes, a feature not available for the linear estimators when p<2.

email: linyuan@math.unh.edu

INFERENCE FOR THE PROPORTIONAL ODDS MODEL WITH A CHANGE-POINT BASED ON A COVARIATE THRESHOLD

Rui Song* , University of Wisconsin–Madison
Michael R. Kosorok, University of Wisconsin–Madison

We consider maximum likelihood estimation of the parameters for the proportional odds model with a change-point under right censoring. The estimator of the change-point is shown to be n-consistent and its asymptotic distribution is established. The estimator of the other regression parameters as well as the baseline hazard is shown to be uniformly consistent for the pseudo-value maximizing the asymptotic limit of the likelihood. Appropriately standardized, the estimator converges weakly to a Gaussian process. When the change-point related parameter is known, the procedure is semiparametric efficient, achieving the semiparametric information bound for all regular parameter components. We use a simulation study to investigate the finite sample and asymptotic behavior of our estimators. The practical utility of the procedure is illustrated on a Primary Biliary Cirrhosis (PBC) dataset.

email: rsong@stat.wisc.edu

ON KERNEL FUNCTION ESTIMATION FOR CENSORED DATA

Kagba N. Suaray*, University of California, Riverside

We study kernel density and survival function estimation for censored data. A brief overview of the history and methodology of the technique are given. A mean squared error derivation, based on elementary principles, of the well studied convolution density estimator is provided. We then proceed to introduce a new density estimator, based on the Kaplan-Meier estimator, and using Abramson's variable bandwidth principle. We give its mean squared error expansion, and investigate the improved bias properties. Finally, we run simulations illustrating the favorable large sample characteristics of the new estimators.

email: ksuaray@stat.ucr.edu

PENALIZED LIKELIHOOD BASED CROSS-VALIDATION METHODS FOR SURVIVAL DATA ANALYSIS

Bin Wang*, University of South Alabama

Semi-parametric estimators can be used to analyze survival data subject to both selection bias and censoring by assuming parametric model of the selection function and no assumption on the distribution of the survival time of the target population. However, sometimes the maximum likelihood estimate of the unknown parameter(s) in the selection function may not be obtained by maximizing a pseudo likelihood based on the survival data. In this study, we propose to estimate by using a penalized likelihood and select the smooth parameter by using cross-validation method. The performance of the new estimator will be demonstrated by simulation results.

email: bwang@jaguar1.usouthal.edu

TIME-VARYING FUNCTIONAL REGRESSION FOR PREDICTING REMAINING LIFETIME DISTRIBUTIONS FROM LONGITUDINAL TRAJECTORIES

Hans-Georg Müller*, University of California, Davis
Ying Zhang, University of California, Davis

A recurring objective in longitudinal studies on aging and longevity has been the investigation of the relationship between age-at-death and current values of a longitudinal covariate trajectory that quantifies reproductive or other behavioral activity. We propose a novel technique for predicting age-at-death distributions for situations where the predictors are entire covariate histories. The predictor trajectories up to current time are represented by time-varying functional principal component scores, which are continuously updated as time progresses and are considered to be time-varying predictor variables that are entered into a class of time-varying functional regression models that we propose. We demonstrate for biodemographic data, for which age-at-death typically is completely observed, how these methods can be applied to obtain estimates of remaining lifetime distributions and predictions for age-at-death. Estimates and predictions are obtained for individual subjects, based on their behavioral trajectories, and include a dimension-reduction step by projection on a single index. We illustrate the proposed techniques with data on longitudinal daily egg- laying for female, predicting remaining lifetime and age- at-death distributions from individual event histories observed up to current time.

email: yzhang@wald.ucdavis.edu

---

8. SURVEY DATA METHODS

CAN POPULATION ESTIMATES WITH BRIDGED-RACE CATEGORIES BE IMPROVED USING THE CENSUS QUALITY SURVEY

Deborah D. Ingram*, National Center for Health Statistics

Census 2000 allowed respondents to report more than one race; most birth and death certificates only permit the reporting of one race. To permit the calculation of race- specific birth and death rates, Census 2000 multiple-race groups are being “bridged” to single-race categories. Logistic and multi-logit models were fit to pooled 1997- 2000 National Health Interview Survey data to predict preferred single-race category for multiple-race respondents. The resulting parameter estimates have been used to obtain county-age-sex-Hispanic origin –specific probabilities of selecting each of the possible single-race categories for the multiple-race groups and these probabilities have been applied to Census 2000 population files to obtain bridged-race population counts. The Census Quality Survey (CQS) used a split panel design to obtain race data using both “Mark 1 or more races” and “Mark 1 race” questions. The resulting multiple-race sample is considerably larger than the NHIS multiple-race sample. The NHIS bridging models have been replicated using the CQS. This presentation examines the impact on the model parameters and on resulting population counts of using additional Census 2000 contextual variables (such as segregation indexes).

email: DDIngram@cdc.gov

AN APPROACH TO ESTIMATING THE DISTRIBUTION AND BIAS OF COMPARATIVE FIT INDICES USING BINARY DATA

Zara E. Sadler*, Medical University of South Carolina
Barbara C. Tilley, Medical University of South Carolina
Philip F. Rust, Medical University of South Carolina
Peng Huang, Medical University of South Carolina
Linda M. Kaste, University of Illinois at Chicago

This research uses an existing dataset from NHANES III to explore the distribution of goodness of fit (GOF) statistics and potential for bias when using bootstrapping techniques to compare Bentler's Comparative fit index (CFI) for independent factor analytic models in the confirmatory setting, when fitting binary data. We focused on the CFI as the bootstrapped distribution of this fit statistic was the most skewed of the three fit statistics (Comparative Fit Index (CFI), Tucker-Lewis fit index (TLI), and the Root Mean Square Error Approximation (RMSEA)) and had the most potential for bias when estimating the variance. We explored the large and small sample properties of the bootstrapped variance estimates for the CFI using proper and improper treatment of the outcome variable. In contrast to the work by Bollen for continuous data, we found minimal bias using Mplus regardless of whether we used large or small samples. We also mis-specified the factor structure and found similar results. Our results suggest for categorical data, analyzed using Mplus, bootstrapping techniques without adjustment for bias can be used to obtain confidence intervals on the CFI. While we suggest that investigators calculate bias when computing desired confidence intervals, we expect Mplus results to be robust to most situations if the model is not mis-specified.

email: sadlerze@musc.edu

FINITE POPULATION CUMULATIVE DISTRIBUTION FUNCTIONS AND MEASUREMENT ERROR

Jeremy Aldworth*, RTI International

If data sampled from a finite population are contaminated with measurement error, then Horvitz-Thompson (HT) estimators of nonlinear estimands are typically biased. The finite population CDF is a nonlinear estimand of particular interest in studies assessing population thresholds. The Stefanski-Bay (SB) estimator uses a SIMEX argument to address this bias, but some approximations in the underlying theory depend on the noise variance being small. Hence, it does not address the bias as well if the variance is large. Two alternative model-based estimators are compared with the HT and SB estimators. These estimators possess good model-based unbiasedness and optimality properties, but the question of real interest is in their corresponding design-based properties. Algebra describing these properties does not lend itself to easy interpretation, but initial simulation studies show promising results. For example, under Gaussian conditions, they have very good design-based unbiasedness properties that do not appear to be affected by the level of noise variance, and their design-based MSE estimates appear to be consistently smaller than the corresponding MSE estimates of the SB estimator.

email: jaldworth@rti.org

MODEL-BASED ESTIMATES OF THE FINITE POPULATION MEAN FOR TWO-STAGE CLUSTER SAMPLES WITH UNIT NONRESPONSE

Ying Yuan*, University of Michigan
Roderick J. A. Little, University of Michigan

We propose new model-based methods for unit nonresponse in two-stage survey samples. In particular, we consider the model-based approach that treats the clusters as random effects. We show that the usual random-effects model estimator of the population mean (RE) is biased in the setting of unit nonresponse unless nonresponse is missing completely at random, which makes the often unrealistic assumption that the response rates are unrelated to cluster characteristics. This fact motivates modifications of RE that allow the cluster means to depend on the response rates in the clusters. Two approaches are considered, one that includes the observed response rate as a cluster-level covariate (RERR), and one based on a probit model for response (NI1). The former approach is simpler than NI1 but approximate, in that uncertainty in estimating the response rates is not taken into account. We show by simulations that estimators from RERR and NI1 can correct the bias of RE, and have comparable or lower root mean squared error than WT in a variety of simulation settings. We also consider another nonignorable model estimate of the mean (NI2) that removes the bias of WT, NI1 and RERR when there is association between response and the survey outcome within the clusters.

email: yuany@umich.edu

UNBALANCED RANKED SET SAMPLING FOR ESTIMATING A POPULATION PROPORTION UNDER IMPERFECT RANKINGS

Haiying Chen*, Wake Forest University School of Medicine
Elizabeth Stasny, The Ohio State University
Douglas Wolfe, The Ohio State University

The application of unbalanced ranked set sampling (RSS) to estimation of a population proportion has been studied for the perfect ranking situation. When the rankings are not perfect, the probabilities of success for the judgment order statistics incorporate information on ranks as well as on ranking errors. The objective of this paper is to investigate the effect of imperfectness in rankings on unbalanced RSS for binary variables and provide methods to obtain estimates for the probabilities of success for the judgment order statistics so that Neyman allocation can be implemented. We also use a substantial data set, the NHANES III data, to demonstrate the feasibility and benefits of Neyman allocation in RSS for binary variables in the case of imperfect rankings.

email: hchen@wfubmc.edu

ESTIMATION OF PREVALENCE OF OVERWEIGHT IN SMALL AREAS-A ROBUST EXTENSION OF THE FAY-HERRIOT MODEL

Dawei Xie*, University of Pennsylvania
Trivellore E. Raghunathan, University of Michigan
James M. Lepkowski, University of Michigan

Hierarchical model such as Fay-Herriot model is often used to develop small area estimates. It might perform well overall but overshrink estimates for some areas with "extreme" direct estimates even though they are based on large samples and are relatively reliable. We propose a robust version of Fay-Herriot model that assumes the area random effects follow a t distribution with a known degree of freedom. Monte Carlo Markov Chain (MCMC) is used to obtain the posterior distribution of small area estimates. The procedure is illustrated in obtaining the prevalence of overweight for the state of Alaska, District of Columbia (DC), and 3112 counties in other 49 states in the United States (US) from the Behavioral Risk Factor Surveillance System (BRFSS) in 2000. The robust model is applied to all the counties, and three strata by county population sizes. The mean square error of small area estimates using different degrees of freedom for stratified or unstratified samples are compared.

email: dxie@cceb.upenn.edu

JACKKNIFE VARIANCE ESTIMATION OF THE REGRESSION AND CALIBRATION ESTIMATOR FOR TWO 2-PHASE SAMPLES

Jong-Min Kim*, University of Minnesota, Morris
Jon E. Anderson, University of Minnesota, Morris

In this paper, we propose a jackknife variance estimator of the population average from two 2-phase samples after imputation. We apply two different sampling methods (Simple Random Sampling and Stratified Random Sampling) to derive jackknife variances of the regression estimator for two samples after imputation under 2-phase sampling. We also apply calibration estimation to ratio imputation in stratified random sampling.

email: jongmink@mrs.umn.edu

HIERARCHICAL BAYESIAN GALERKIN-BASED PARAMETERIZATIONS OF SPATIO-TEMPORAL DYNAMICAL MODELS WITH APPLICATION TO ECOLOGICAL PROCESSES

Ali Arab*, University of Missouri–Columbia
Christopher K. Wikle, University of Missouri–Columbia

Ecological processes are typically very complex and exhibit complicated behavior through many different scales of spatial and temporal variability. Furthermore, ecological analysis problems often include complicated boundaries. This suggests that the incorporation of scientific knowledge (e.g., differential equations) in the statistical model is essential. We consider efficient parameterizations of spatio-temporal models for such processes using Galerkin-based procedures. As an example, we will consider finite element representations to modeling invasive species using reaction-diffusion models with spatially-varying parameters.

email: aa5vf@mizzou.edu

DIAGNOSTIC APPROACHES TO STATISTICAL MODELS INCORPORATING DYNAMIC BIOLOGICAL COMPONENTS

Michael B. Brimacombe*, New Jersey Medical School–UMDNJ

Many biological and growth related processes can be expressed in terms of nonlinear dynamic functions. Often these are based upon population dynamics or specific properties of components evolving through time. In settings where these functions become components in statistical models, usually in a linear model or regression format, such mathematical representations can easily yield likelihood functions that are unstable locally and limit the relevance of statistical estimation and testing. Here the use of new diagnostics such as box-counting along directional profiles of the likelihood function is examined as a measure of likelihood function instability. Examples drawn from population dynamics and growth models are given.

email: brimacmb@umdnj.edu

A HIERARCHICAL BAYESIAN APPROACH FOR DESCRIBING THE SPATIO-TEMPORAL DYNAMICS OF INVASIVE SPECIES

Mevin B. Hooten*, University of Missouri–Columbia
Christopher K. Wikle, University of Missouri–Columbia

Spatial growth and dispersal of biotic organisms as a function of time has been recognized as an important subject throughout the relatively short history of ecology as a science. Ecologists are able to accurately describe survival and fecundity in plant and animal populations, however the dynamics of dispersal are rarely measured and not well understood. Of particular interest are the dynamics of invasive species. Such nonindigenous animals (and plants) can levy significant impacts on native biotic communities. Generally, a successful invasion depends on a sequence of stages, i.e., introduction, establishment, range expansion, and saturation. A better quantitative understanding of these stages would be beneficial to all branches of ecology (and other biological sciences as well). We adopt a hierarchical Bayesian framework for modeling the invasion of such species. Our approach responsibly accounts for various sources of uncertainty as well as spatial and temporally varying detection probabilities which helps alleviate inflated abundance estimates. The dynamics between discrete time points are modeled in a sequence of growth and dispersal stages that allow for birth and mortality as well as dispersal of living organisms among locations.

email: hooten@stat.missouri.edu

TEST FOR INDEPENDENCE BETWEEN MARKS AND POINTS OF A MARKED POINT PROCESS

Yongtao Guan*, University of Miami

A convenient assumption while modeling a marked point process is that the observations (i.e., marks) and the locations (i.e., points) are independent. We propose new graphical and formal testing approaches to test for this assumption. The proposed graphical procedures are easy to obtain and can be used to diagnose the nature and range of dependence between marks and points. The formal testing procedures require only minimal conditions on marks and thus can be applied to a variety of settings. We illustrate these procedures through a simulation study and an application to some real data.

email: yguan@miami.edu

COMBINING INFORMATION FROM MULTIPLE SOURCES TO ESTIMATE THE PROBABILITY OF A RARE EVENT

Philip M. Dixon*, Iowa State University

Successful capture of wasps by the insectivorous cobra lily plant is rare. A total of 157 wasp visits and 2 captures were seen during 376.5 plant hours of direct observation. The per-visit capture probability is 1.3%, but the relative uncertainty in this estimate is large. Increasing the precision by additional direct observation is laborious. Aggregated data on the total number of captures over a defined time interval are easy to collect but provide no information on the number of wasp visits. I develop models to estimate visitation rate (visits per unit time) and per-visit capture probability by combining detailed and aggregated data. Simple models assume constant rates; other models allow rates to be functions of plant characteristics. By evaluating asymptotic relative efficiency, I show that combining data increases the precision of the estimated visitation rate when the capture probability is large. When the capture probability is small, combining data increases the precision of the estimated per-visit capture rate. The increase in precision can be substantial. For the cobra lily data, combining data provides the same precision as over 1700 plant-hours of direct observation.

email: pdixon@iastate.edu

TESTS FOR ORDER-RESTRICTIONS IN ORDINAL DATA: A GRAPHICAL APPROACH

Eric R. Teoh*, University of North Carolina at Chapel Hill
Abraham Nyska, National Institute of Environmental Health Sciences
Uri Wormser, Hebrew University of Jerusalem
Shyamal D. Peddada, National Institute of Environmental Health Sciences

We present a general methodology for testing equality of the responses of several treatment groups, measured on an ordinal scale (such as no, low, medium, high response), against any given ordered-alternative. The proposed test statistic is based on the estimation procedure developed in Hwang and Peddada (1994). It utilizes the Pooled Adjacent Violator Algorithm (PAVA) for scalar parameters subject to a simple-order. We illustrate this procedure by applying it to a dataset from a toxicological study. Simulations we performed suggest that the proposed method maintains an appropriate type I error rate and is at least as powerful as some existing methods. The proposed method also handles a broad class of alternatives and is straightforward to implement.

email: teoh@unc.edu

SOME CHALLENGES ENCOUNTERED WHEN ANALYZING BIOMARKER DATA

Stephen W. Looney*, Louisiana State University Health Sciences Center School of Public Health
Joseph L. Hagan, Louisiana State University Health Sciences Center School of Public Health

In this presentation, we provide a description of some challenges that may be encountered in the analysis of biomarker data, and offer recommendations on how best to deal with these challenges. The issues that we consider are illustrated with examples taken from the biomarker literature and include: assessment of distributional assumptions, measurement of association between predictor and outcome, analysis of cross-classified categorical data, comparison of mean levels across groups, use of correlation coefficients, and statistical comparison of biomarkers. Our recommendations are based on published statistical evidence and we indicate how widely available statistical software can be used to carry out the recommended analyses.

email: sloon1@lsuhsc.edu

10. SURVIVAL ANALYSIS I

ESTIMATING SURVIVAL DISTRIBUTIONS IN THE PRESENCE OF INFORMATIVE CENSORING VIA A LATENT SURVIVAL TIME IMPUTATION APPROACH

Pai-Lien Chen*, Family Health International
Bosny Pierre-Louis, Family Health International and University of North Carolina at Chapel Hill

Censored outcomes due to early discontinuation and loss to follow-up of study participants is one of the major challenges in analyzing time-to-event data. Standard survival distribution estimations are biased when the censoring mechanism is informative. We propose a latent survival time imputation approach to address this problem. Censored data are imputed using mean residual survival times based on Kaplan-Meier estimates or, when auxiliary variables are available, using weighted Kaplan- Meier estimates and Cox regression models. A simulation study demonstrates that the proposed approach reduces the bias of the survival distribution estimations compared with estimations that ignore the informative censoring mechanism. We apply the proposed method to a clinical trial designed to estimate the effect of spermicidal agents on pregnancy outcomes, where loss to follow-up is assumed informative.

email: pchen@fhi.org

VARIABLE SELECTION FOR CENSORED DATA

Brent A. Johnson*, University of North Carolina at Chapel Hill

Variable selection is an important topic in the statistical sciences with broad applications to many substantive areas. Additional difficulty is caused by censoring, which occurs when subjects are not followed until the endpoint of interest has occured so their failure times are unknown. We propose a family of methods based on the accelerated failure time model and Buckley-James estimator. The large and small sample properties of such methods will be considered.

email: brentj@email.unc.edu

A SIMPLE APPROACH TO THE ESTIMATION OF THE SURVIVAL FUNCTION BASED ON TWO-STAGE SAMPLING FOR THE DEPENDENT CENSORSHIP

Seungyeoun Lee*, Sejong University–Seoul, Korea

A simple approach is proposed for estimating the survivor function under the dependent censorship. This approach is based on the two-stage sampling design proposed by Lee and Wolfe (1998), which involves further follow-up of a subset of lost-to-follow-up censored subjects. Under the two-stage sampling design, a proportional hazards estimator and a semi-Markov estimator were proposed for the dependent censoring model by Lee and Wolfe (1998) and Lee and Tsai(2004), respectively. In this paper, we propose a new estimator for the survivor function under a non-homogeneous Markov model. The consistency of the proposed estimator is derived and the estimation procedure is illustrated with an example of lung cancer clinical trial. Finally simulation studies are performed to compare the proposed estimator with other estimators under the proportional hazards alternative and non-proportional hazards alternative, respectively.

email: leesy@sejong.ac.kr

NONPARAMETRIC ESTIMATION OF THE CONCORDANCE CORRELATION COEFFICIENT UNDER UNIVARIATE CENSORING

Ying Guo*, Rollins School of Public Health, Emory University
Amita K. Manatunga, Rollins School of Public Health, Emory University

This paper proposes a nonparametric estimator of Lin's (1989, Biometrics, 255-268) concordance correlation coefficient (CCC) for assessing agreement of multivariate continuous survival times. A time-dependent CCC is also developed for measuring the agreement among survivors at each time point. In the presence of censoring with survival outcomes, we propose to estimate Lin's CCC and the time-dependent CCC through Lin and Ying's (1993, Biometrika, 573-581) nonparametric estimator of the bivariate survival function under univariate censoring. The presented estimators are proven to be strongly consistent and asymptotically normal, with consistent bootstrap variance estimators. Numerical studies are performed to evaluate the accuracy of the estimators and their bootstrap variance estimators. Finally, using the proposed methods, the results from a prostate cancer study are presented where time to cancer recurrence is measured by two different definitions.

email: yguo2@sph.emory.edu

CONSTRUCTING EXACT CONFIDENCE BOUNDS FOR THE TRUE SURVIVAL CURVE USING THE KAPLAN-MEIER SURVIVAL FUNCTION

Craig B. Borkowf*, Centers for Disease Control and Prevention

We present an exact method for constructing confidence bounds for the true survival curve as estimated by the Kaplan-Meier survival function (KMSF). We describe the properties that the exact confidence bounds should satisfy, as a consequence of the definition of the KMSF. In particular, we argue that the exact upper bound should decrease only when events occur, but the exact lower bound should decrease when either events or censorings occur. For the proposed exact method, the width of the exact confidence interval at any given time is proportional to the current weights of the observations still present in the sample under Efron's redistribute-to-the-right algorithm. We present the results of a simulation study to show that the proposed exact method gives at least nominal coverage in a variety of situations, even for small sample sizes, although coverage can be quite supranominal. We also compare the performance of this exact method with that of several traditional approaches for confidence interval construction. Finally, an application to a data set from a leukemia clinical trial illustrates the practical importance of using exact confidence bounds to make inference to the underlying survival distribution.

email: CBorkowf@cdc.gov

TESTING GOODNESS-OF-FIT OF A TRUNCATION MODEL

Micha Mandel*, Harvard School of Public Health
Rebecca Bentensky, Harvard School of Public Health

Several goodness-of-fit tests of a lifetime distribution have been suggested in the literature; many take into account censoring and/or truncation of the event times. In some contexts, a goodness-of-fit test for the truncation distribution is of interest. In particular, knowledge of the truncation distribution can be exploited to derive more efficient estimates of the lifetime distribution. Also, in cross-sectional sampling, the truncation events signify the incidence rate of some event and it is of interest to test for constancy of that rate. The duality of lifetime and truncation in the absence of censoring enables methods for testing goodness-of-fit of the lifetime distribution to be used for testing goodness-of-fit of the truncation distribution. However, under random censoring, this duality does not hold and different tests are required. We develop several goodness-of-fit tests for the truncation distribution and investigate their performance in the presence of censored event times.

email: mmandel@hsph.harvard.edu

A PEARSON GOODNESS-OF-FIT TEST FOR INTERVAL CENSORED DATA

Denise Babineau*, University of Waterloo
Jerry Lawless, University of Waterloo

At this time, there are very few methods to assess the fit of a hypothesized parametric model for general interval censored data when the alternative hypothesis is unspecified. A comparison of expected and observed failures, commonly seen in Pearson tests for grouped data, is impossible because each subject's failure interval may overlap with other subjects' failure intervals. This talk introduces a Pearson test to overcome this problem by using an alternative multinomial structure as the distribution for the number of failures in each interval. The distribution and power of this test is also discussed.

email: dbabinea@math.uwaterloo.ca

PREDICTION ERROR ESTIMATION: A COMPARISON OF RESAMPLING METHODS

Annette M. Molinaro*, National Cancer Institute
Ruth Pfeiffer, National Cancer Institute
Richard Simon, National Cancer Institute

High dimensional genomic measurements are a relatively new type of data that give rise to two types of problems: class discovery and class prediction. In class prediction, one uses training data, containing both the class membership Y and p covariates X to define a classification algorithm to predict class membership of future observations based on X alone. It is inherently important to correctly assess the prediction error of a given model. In the absence of independent validation data, there are numerous techniques for assessing prediction error by implementing some form of partitioning or resampling of the original data. Each technique involves dividing the data into a learning set and a test set. Depending on the resampling method, the learning set may further be divided into a training set and an evaluation set. These methods range in complexity from a test split to v-fold cross-validation to Monte-Carlo v-fold cross-validation to bootstrap variants. To date, there has not been a comprehensive comparison of these resampling methods for prediction error estimation. We exhaustively compare performance of the resampling methods using simulations, proteomic and microarray data encompassing increasing sample sizes with high complexity. The results elucidate the 'best' resampling technique to guide future studies.

email: molinaran@mail.nih.gov

BAYESIAN IDENTIFICATION OF PROGNOSTIC MOLECULAR SIGNATURES FOR SURVIVAL PHENOTYPES

Dabao Zhang*, University of Rochester Medical Center

The high-throughput biotechnologies usually provide enormous molecular features containing the genomic and/or proteomic information for each of limited number of individuals in order to understand the genetic basis of important clinical outcomes. This variable selection issue is challenging because of the ``large p small n' data, especially when the clinical outcomes are survival times. With a diffuse gamma process prior on the baseline hazard function as justified by Kalbfleisch (1978), we propose a Bayesian variable selection approach to statistically identify molecular signatures for survival clinical outcomes with ``large p small n' data. This approach is implemented via Gibbs sampling scheme and Laplace approximation. Simulation study shows that it performs much better than the existing approaches.

email: dabao_zhang@urmc.rochester.edu

USING LONGITUDINAL GENOMIC DATA TO PREDICT FAILURE OUTCOMES

Natasa Rajicic*, Harvard University
Dianne Finkelstein, Harvard University
David Schoenfeld, Harvard University

We describe an approach to the survival analysis of longitudinally collected genomic data. We construct a measure of association between the survival endpoint and gene expressions collected over time and find significance levels using permutations. This nonparametric approach does not depend on any unverifiable assumptions about the unknown distributions of gene expressions. The issue of high dimensionality and dependence present in the genomic data is addressed through a multiple testing procedure. We also address missing data problem which occurs as a result of using permutations on possibly censored, longitudinal data. Our proposed method is illustrated on a dataset from a multi-center research study of inflammation and the host response to injury that aims to uncover the biological reasons why patients can have dramatically different outcomes after suffering a traumatic injury.

email: nrajicic@hsph.harvard.edu

STATISTICAL MODELS FOR CHARACTERIZING FUNDAMENTAL PATTERNS UNDERLYING GENE EXPRESSION PROFILES

Fei Long*, University of Florida
Tian Liu, University of Florida
Rongling Wu, University of Florida

Recent development of DNA microarray technology that enables the production of massive amounts of genomic data has highlighted the need for powerful pattern recognition techniques that can discover biologically meaningful knowledge in large datasets. In particular, the identification of fundamental patterns or clusters of genes for temporal profiles of their expression can provide a quantitative and testable framework for cutting edge research between gene action and development. Here, we develop a novel statistical model for clustering gene expression profiles based on their underlying physiological functions. This model integrates the Fourier series approximation of time-dependent gene expression with the statistical modelling of the structure of the covariance matrix across the time course within the framework of finite mixture models. By estimating and testing the Fourier coefficients that determine the shapes of temporal profiles, the patterns of gene expression can be compared and their functions determined. The statistical properties of our model are studied through computer simulation.

email: lfei@ufl.edu

FEATURE-SPECIFIC CONSTRAINED LATENT CLASS ANALYSIS FOR GENOMIC DATA

Andres Houseman*, Harvard University
Brent A. Coull, Harvard University
Rebecca A. Betensky, Harvard University

Genomic data are typically characterized by a moderate to large number of categorical outcomes observed for relatively few subjects. Some of the outcomes may be missing or noninformative. An example of such data is Loss of Heterozygosity (LOH), a dichotomous outcome, observed on a moderate number of genetic markers. We first present a latent class model where, conditional on (unobserved) membership in one of k classes, the outcomes are independent with probabilities determined by a regression model of low dimension p. Using a family of penalties including the Ridge and Lasso, we extend this model to address higher dimensional problems. Finally, we present an orthogonal map that transforms marker-space to a space of 'features' for which the constrained model is better behaved. We demonstrate these methods on LOH data collected at 19 markers from 93 brain tumor patients. Additionally, we show that posterior classes obtained from this method can predict survival.

email: ahousema@hsph.harvard.edu

SHARP SIMULTANEOUS INTERVALS FOR THE MEANS OF SELECTED POPULATIONS WITH APPLICATION TO MICROARRAY DATA ANALYSIS

Jing Qiu*, University of Missouri–Columbia
Gene J. T. Hwang, Cornell University

Simultaneous inference is a challenge when the number of populations, N,or the dimensionality is large. In some situations including microarray experiments, the scientists are only interested in the K populations with parameters (such as means) that have the most extreme estimates. In these cases, can we construct simultaneous intervals for the means corresponding to these K selected populations? The answer is yes as demonstrated here and the approach allows us to cut down the dimensionality of the problem from N to K. The naive simultaneous intervals for the K means (applied directly without taking into account the selection) have low coverage probabilities. We take an Empirical Bayes approach (or an approach based on the mixed effect model) and we construct simultaneous intervals with good coverage probabilities. For N=10,000 and K=100, typical for microarray data, the lengths of our intervals could be 82\% shorter than those of the Bonferroni's N-dimensional simultaneous intervals and 77\% shorter than those of the naive K-dimensional simultaneous intervals.

email: qiujing@missouri.edu

12. RECENT ADVANCES IN THE ANALYSIS OF RECURRENT EVENTS DATA

RECURRENT EVENTS AND LONGITUDINAL MARKERS

Edsel A. Pena*, University of South Carolina
Elizabeth H. Slate, University of South Carolina
Jun Han, University of South Carolina

This talk will deal will recent developments in the modelling and analysis of recurrent events and of longitudinal markers. We will first consider a general class of recurrent event models and the statistical procedures for this class of models. Next, we will then discuss an extension of this class of models to incorporate longitudinal markers which could potentially improve our knowledge regarding the recurrent event process. We discuss varied aspects for this class of models such as parameter interpretation, parameter estimation, and issues regarding the gain achieved by utilizing longitudinal markers through the proposed class of models. Inference methods for this class of models will be presented. This joint class of models will have applicability in a variety of settings, notably in biomedical and public health situations. As such, some real applications will be illustrated.

email: pena@stat.sc.edu

SEMIPARAMETRIC INFERENCE OF SUCCESSIVE DURATIONS

Yijian Huang*, Emory University

For many chronic diseases, a bi-state progressive process provides a useful model for the disease progression before reaching death. For example, after mastectomy a breast cancer patient progresses through disease-free and relapse states. Often, scientific interests lie in the successive durations. For the one-sample problem with incomplete follow-up data, recent investigations have focused on nonparametric inference. However, in many practical situations, the distribution of the second duration is nonparametrically nowhere identifiable. Furthermore, most existing approaches require a rather restrictive censoring mechanism and have difficulty in predicting the process with given history. To address these issues, we suggest a copula model for the association between the two durations, while leaving the marginals unspecified. This semiparametric model ensures marginal distribution of the second duration to be identifiable except for the tail. An inference procedure is proposed and illustrated with a clinical study.

email: yhuang5@emory.edu

FRAILTY IN THE ACCELERATED GAP TIMES MODEL

Robert L. Strawderman*, Cornell University

A natural choice of time scale for analyzing recurrent event data is the gap, or sojourn, time between successive events. It is often reasonable to assume correlation exists between the sojourn times within a subject. In this talk we extend the accelerated failure time model to the case of dependent recurrent event data by first constructing an appropriate intensity model for the case of independent gap times and then incorporating frailty. Estimators are motivated using semiparametric efficiency theory and lead to interesting generalizations of the class of rank statistics typically used for the accelerated failure time model. The proposed methodology is illustrated by simulation and data analysis. Extensions to more complex forms of dependence are also considered.

email: rls54@cornell.edu

ANALYZING RECURRENT LONGITUDINAL DATA WITH COMPLICATION OF INFORMATIVE CENSORING

Mei-Cheng Wang*, Johns Hopkins University

Longitudinal data are usually defined as repeated measurement data where sampling points are independent of the measurements. This talk considers analysis of repeated measurement data in the situation that sampling points are possibly recurrent events, and repeated measurements are possibly correlated with recurrent events. The outcome measures of interest include both the recurrent events and the repeated measurements measured at times of events. Suppose the observation of repeated measurements is terminated either by censoring or occurrence of a failure event, whichever occurs first. A mean function (MF) at t is defined as the expectation of the total measurements in unit time at t. This talk will introduce nonparametric and semiparametric models for the described data, and develop statistical methods for estimating the cumulative MF and parameters in a MF regression model. Estimation of other related functions will be briefly discussed.

email: mcwang@jhsph.edu

13. NATURAL RESOURCE ESTIMATION FOR SMALL AREAS

NONPARAMETRIC SMALL AREA ESTIMATION USING PENALIZED SPLINE REGRESSION

Jean Opsomer*, Iowa State University
Jay Breid, Colorado State University
Gerda Claeskens, Katholieke Universiteit Leuven (Belgium)
Goeran Kauermann, Universitaet Bielefeld (Germany)
Giovanna Ranalli, Universita di Perugia (Italy)

Penalized spline regression provides a convenient framework for constructing nonparametric small area estimators in natural resource surveys. At the population level, the relationship between the variable of interest and continuous auxiliary variables is modelled as a smooth but otherwise unspecified function, while the small area effects are incorporated in the model through a random effects specification. We present the resulting nonparametric small area estimator and discuss its statistical properties. The methodology is applied to a survey of lakes in the Northeastern US.

email: jopsomer@iastate.edu

SMALL AREA ESTIMATION OF FOREST ATTRIBUTES FOR A TEMPORALLY CONTINUOUS SAMPLING DESIGN

Francis A. Roesch*, USDA Forest Service

Mixed estimation is evaluated for estimating the current and dynamic states of small areas sampled by a USDA Forest Service annual forest sampling design. A globally defined mixed estimator is applied and refined at a series of increasingly localized scales in a simulation in order to determine the optimum information selection areas for various variables of interest.

email: froesch@fs.fed.us

MAP-BASED ESTIMATION OF FOREST AREA

Ronald E. McRoberts*, USDA Forest Service

The Forest Inventory and Analysis (FIA) program of the USDA Forest Service collects a suite of ecological and mensurational observations for an array of field plots systematically distributed across the nation at the intensity of approximately one plot per approximately 2,400 hectares (6,000 acres). The exact locations of plots remain confidential to protect the ecological integrity of the sample, to protect the proprietary nature of some plot information, and to protect land owners from unwelcome intrusions. Nevertheless, users of FIA data frequently desire estimates for arbitrarily defined areas of interest (AOI) which would normally require plot locations. An alternative to providing plot locations is to use the plot data and satellite imagery to produce a map of the forest attribute of interest, and then calculate area estimates by summing or averaging over the map-based, pixel-level predictions of the attribute of interest. Data from inventory plots and images for three dates of Landsat Thematic Mapper satellite imagery were used with a k- Nearest Neighbors approach to predict percent canopy cover for each 30m x 30m pixel. Estimates of the means and standard errors of map-based percent forest canopy cover were calculated for circular areas with radii as small as a few kilometers and compared to estimates based on the traditional design-based estimation approach.

email: rmcroberts@fs.fed.us

ESTIMATING COHO SALMON ABUNDANCE IN SMALL STREAM BASINS USING A SPACE-TIME MODEL WITH COVARIATES

Don L. Stevens Jr.*, Oregon State University
Ruben A. Smith, Oregon State University

The Oregon Department of Fish & Wildlife conducts annual surveys of spawning coho salmon in Oregon coastal streams. A rotating panel design has been in place since 1998, and some 50 years of index sample information is available. Companion probability surveys of juvenile abundance and habitat condition are also conducted. The survey was designed to estimate population status at the level of five monitoring areas (MA's), which were originally based on Evolutionarily Significant Units. However, recent developments have indicated that the five MAs comprise 30+ more or less independent biological populations for which estimates of population size and viability are needed. In addition, landscape information, non-probability habitat condition, and a measure of intrinsic potential are available. We use this information to construct a space-time auto-regressive moving average model to estimate spawner abundance for stream basins with few sample sites.

email: stevens@science.oregonstate.edu

14. DEALING WITH MALDI-TOF/SELDI-TOF PROTEOMIC DATA: EXPERIMENTAL DESIGN AND LOW-LEVEL PROCESSING

MALDI-TOF/SELDI-TOF: BACKGROUND AND DESIGN ISSUES

Keith A. Baggerly*, University of Texas M. D. Anderson Cancer Center
Jeffrey S. Morris, University of Texas M. D. Anderson Cancer Center

This talk is in two parts. The first part of this talk provides a common background for all talks in this session.

I. HOW MALDI/SELDI-TOF WORKS.
We provide an introduction to the basic mechanics of MALDI-TOF and SELDI-TOF mass spectrometry: how a sample is processed to produce a spectrum with peaks corresponding to proteins.

II. SIGNAL IN NOISE: BIAS OR BIOLOGY?
We introduce some issues of experimental design in the context of case study counterexamples, illustrating how various findings that might be attributed to biology can equally well be explained by experimental biases. These biases can be seen in the data directly, raising issues of exploratory data analysis. These findings have caused some debate as to whether the technology is ready for 'prime time'; we also provide some background on the current regulatory context.

email: kabagg@mdanderson.org

ENHANCEMENT OF SENSITIVITY AND RESOLUTION OF SURFACE-ENHANCED LASER DESORPTION/ IONIZATION TIME-OF-FLIGHT MASS SPECTROMETRIC RECORDS

Dariya I. Malyarenko*, College of William and Mary
William E. Cooke, College of William and Mary
Eugene R. Tracy, College of William and Mary
Michael W. Trosset, College of William and Mary
Haijian Chen, College of William and Mary
Dennis M. Manos, College of William and Mary
Maciek Sasinowski, INCOGEN, Inc.
John Semmes, Eastern Virginia Medical School
Gunjan Malik, Eastern Virginia Medical School

SELDI-TOF instruments typically have low mass resolution and considerable electronic noise associated with their detectors. The net result is unnecessary overlapping of peaks, apparent mass jitter, and difficulty in distinguishing mass peaks from the background noise. Many of these effects can be reduced by using standard time series signal processing techniques to improve the background subtraction, to reduce noise and to enhance the mass calibration. Moreover, time series filtering can also produce higher resolution mass spectra. We will illustrate these enhancements as applied to a clinical data set.

email: cooke@physics.wm.edu

DENOISING OF MASS SPECTROMETRY DATA: WAVELETS AND AVERAGING

Jianhua Hu*, University of Texas M. D. Anderson Cancer Center
Kevin R. Coombes, University of Texas M. D. Anderson Cancer Center
Keith A. Baggerly, University of Texas M. D. Anderson Cancer Center
Jeffrey S. Morris, University of Texas M. D. Anderson Cancer Center

This talk will focus on some approaches of denoising mass spectrometry data that has been produced according to a reasonable design. The issues addressed involve decomposing the data into three components: a low-frequency baseline, additive white noise, and a set of peaks. The peaks are what we are interested in, however, recognizing them is difficult due to the fact that the shape of a peak can depend on its position in the spectra. A wavelet approach is used to capture this multi-scale nature of the data. The approach can be further enhanced by a simple trick not yet implemented in standard software: averaging spectra within a group.

email: jhu@mdanderson.org

---

15. AT THE CROSSROADS OF BIOSTATISTICS AND SECURITY BIOMETRICS

(SECURITY) BIOMETRICS FROM A (STATISTICAL) BIOMETRICS PERSPECTIVE

Michael E. Schuckers*, St. Lawrence University and Center for Identification Technology Research

Devices such as fingerprint scanners and iris recognition systems are becoming increasingly prevalent. They have already or will soon appear on passports, at border crossings and in airports. In this talk, I describe the nature of these devices and the places where statistics and biostatistics can play an important role in this vital area. I begin by describing how these devices collect data, process that data and make a decision to accept or reject an individual. This process is basically a classification problem for deciding between two groups: permitted users and non-permitted users. (In bio-authentication, these groups are referred to as genuine users and imposter users.) I will also discuss estimation of false accept and false reject rates using repeated binary measures as is typical in the testing of these devices. A discussion of issues related to experimental design for matching performance evaluation will also be provided. Finally, I will provide suggestions for ways that statisticians and bio-statisticians can become involved in bio-authentication.

email: schuckers@stlawu.edu

STATISTICAL ISSUES IN BIOMETRIC IDENTIFICATION

David Banks*, Duke University

Biometric identification is emerging as a key component of national security planning. But there are hard statistical issues---how to estimate the false alarm rates, how to estimate the missed alarm rates, how to combine information from multiple matching systems, and how to develop fast algorithms. This talk surveys some of the work that has emerged, largely from a data mining perspective, and points out related methodology in text retrieval systems and constrained nonparametric function estimation in the context of multidimensional scaling.

email: banks@stat.duke.edu

THE DNA TESTING EXPERIENCE

Bruce S. Weir*, North Carolina State University

The current activity in 'biometric' methods for identification or individualization in the security setting has some parallels to activity ten years ago when DNA profiling for forensic purposes was being established. From an ENAR perspective, the forensic DNA experience showed very clearly that sound statistical methodology is essential. Also important is the fact that biometric markers, including DNA profiles, have a biological basis and so are shaped by evolutionary forces and require that attention be paid to relationships among people in the same family or in the same population. It is also appropriate to recall the discussion that accompanied the introduction of fingerprints in the early 1900s and how that methodology supplanted earlier methods based on anthropometric categories. DNA profiles have achieved the same level of general acceptance accorded to fingerprints, but care will be needed for general acceptance of biometric identification.

email: weir@stat.ncsu.edu

PREDICTION BY SUPERVISED PRINCIPAL COMPONENTS

Brad Efron, Stanford University
Trevor Hastie, Stanford University
Ian Johnstone, Stanford University
Rob Tibshirani*, Stanford University

Least Angle Regression (LARS) is a new model selection algorithm. It is a useful and less greedy version of traditional forward selection methods. Three main properties of LARS are derived. (1) A simple modification of the LARS algorithm implements the Lasso, an attractive alternative to OLS that constrains the sum of the absolute regression coefficients. The LARS modification calculates all possible Lasso estimates for a given problem in an order of magnitude less computer time than previous methods. (2) A different LARS modification efficiently implements epsilon Forward Stagewise linear regression, another promising new model selection method; this connection explains the similar numerical results previously observed for the Lasso and Stagewise, and helps understand the properties of both methods, which are seen as constrained versions of the simpler LARS algorithm. (3) A simple approximation for the degrees of freedom of a LARS estimate is available, from which we derive a Cp estimate of prediction error; this allows a principled choice among the range of possible LARS estimates. There are strong connections between the epsilon forward stagewise regression and the boosting technique popular in machine learning. These connections offer new explanations for the success of boosting.

email: tibs@stat.stanford.edu

THE ENTIRE REGULARIZATION PATH FOR THE SUPPORT VECTOR MACHINE

Trevor Hastie*, Stanford University
Saharon Rosset, Stanford University
Rob Tibshirani, Stanford University
Ji Zhu, Stanford University

The Support Vector Machine is a widely used tool for classification. Many efficient implementations exist for fitting a two-class SVM model. The user has to supply values for the tuning parameters: the regularization cost parameter, and the kernel parameters. It seems a common practice is to use a default value for the cost parameter, often leading to the least restrictive model. In this paper we argue that the choice of the cost parameter can be critical. We then derive an algorithm that can fit the entire path of SVM solutions for every value of the cost parameter, with essentially the same computational cost as fitting one SVM model. We illustrate our algorithm on some examples, and use our representation to give further insight into the range of SVM solutions.

email: hastie@stanford.edu

BOOTSTRAP CONFIDENCE BANDS

Laura L. Johnson*, National Institutes of Health
Paula Diehr, University of Washington

In order to compare non-monotonic trajectories over time, Bootstrap Percentile Confidence Bands are used to allow comparisons between the two non-monotonic curves while obtaining the correct coverage. Special emphasis is placed on sampling from a longitudinal dataset. An example is drawn from the authors' proposed way to jointly model longitudinal health status data and survival data, the Probability of being Alive and Healthy (PAH), while avoiding imputing data after death.

email: lauralee.johnson@comcast.net

SIMULATION OF BREAST CANCER SCREENING AMONG WOMEN VETERANS

Wenyaw Chan*, University of Texas Health Science Center at Houston School of Public Health
David R. Lairson, University of Texas Health Science Center at Houston School of Public Health
David P. Smith, University of Texas Health Science Center at Houston School of Public Health
Yen-Peng Li, University of Texas Health Science Center at Houston School of Public Health

This research is to present a stochastic model and to perform the simulation study for comparing the life expectancy of women veterans studied by the Project H.O.M.E. Project H.O.M.E. is a four year study of breast cancer screening and prevention. The proposed model of this research includes the natural history of breast cancer characterized as pre-clinical, local, regional and distant stages. Participants were encouraged to take the mammography screen every other year following the guidelines by Project H.O.M.E. For the control group, and various intervention groups, the compliance rates were estimated. For the year that the mammography screen is not due, a self-examination is assumed. Therefore, each subject still has a probability of detecting cancer by herself. For the mammography screening, sensitivity, false-positive and false negative were both included in the model. A medical intervention was assumed to follow any cancer participant. The life expectancy was simulated for these patients, assuming it follows an exponential distribution. This research will also calculate the variance of the outcome variable (lifetime of the subject) and hence estimate the number of simulation runs needed for a fixed range of outcome estimation.

email: Wenyaw.Chan@uth.tmc.edu

ESTIMATING INCREMENTAL COST-EFFECTIVENESS RATIOS AND THEIR CONFIDENCE INTERVALS WITH DIFFERENTIALLY CENSORED DATA

Hongkun Wang*, University of Rochester
Hongwei Zhao, University of Rochester

With medical cost escalating over recent years, cost analysis is being conducted more and more to assess economical impact of new treatment options. An incremental cost-effectiveness ratio is a measure that assesses the additional cost for a new treatment for saving one year of life. In this paper, we consider cost effective analysis for new treatments evaluated in a randomized clinical trial setting with staggered entries. In particular, the censoring times are different for cost and survival data. We propose a method for estimating the incremental cost-effectiveness ratio and obtaining its confidence interval when differential censoring exists. Simulation experiments are conducted to evaluate our proposed method. We also apply our methods to a clinical trial example comparing the cost-effectiveness of implanted defibrillators with conventional therapy for individuals with reduced left ventricular function after myocardial infarction.

email: wang@bst.rochester.edu

MODELING INFANT BIRTHWEIGHT

Katherine J. Hoggatt*, University of California, Los Angeles
Sander Greenland, University of California, Los Angeles
Beate R. Ritz, University of California, Los Angeles

Low weight at birth has been associated with both poor infant health and adult-onset chronic disease. Although there is debate about which aspects of the population birthweight distribution are relevant to epidemiologic research, most investigators have focused on either average birthweight or the probability of low birthweight (<2500g), modeled using linear or logistic regression, respectively. Both these classes of parametric models make strong assumptions about the relation between birthweight and its predictors. We found that even qualitative inferences about whether a factor is harmful or protective can be sensitive to the choice of model. We propose an alternative method for analyzing birthweight using semi-parametric models developed for survival analysis.

email: khoggatt@ucla.edu

ESTIMATING THE EFFECTIVENESS OF MENTAL HEALTH SERVICES IN A STUDY OF THE 1998 U.S. EMBASSY BOMBING IN NAIROBI, KENYA

Haekyung Jeon-Slaughter*, University of Oklahoma Health Sciences Center
Betty Pfefferbaum, University of Oklahoma Health Sciences Center
Carol S. North, Washington University School of Medicine
Pushpa Narayanan, University of Oklahoma Health Sciences Center
Lee Ann Ross, United States Agency for International Development, Kenya

ANOVA and ANCOVA are commonly used in mental health research to analyze data on symptom outcomes. Both approaches assume either a random sampling or normal distribution of outcomes, rare in mental health research. Thus, these approaches may yield inconsistent estimation of outcome when used with censored data or with databases characterized by structural problems. This study proposes a model combining two approaches, censored regression and repeated measures analysis, to analyze data from a study of mental health reactions to the 1998 bombing of the U.S. Embassy in Nairobi, Kenya. The study was conducted three years after the bombing with a volunteer sample of 206 adults receiving mental health services and with 24 clinicians who provided those services. Participants were invited to join the study after a mental health screening. Self- and clinician-reported symptom data were collected at the initiation of services and again after six sessions. These data are left censored because the screening eliminated those with few or no symptoms. In addition, structural problems include unequal numbers of participants per clinician and different lengths of time between initiating and completing the six sessions. GEE will be used to estimate the model parameters and to conduct simulation studies to compare estimates of symptom change using ANOVA, ANCOVA and our proposed model.

email: hattie-jeon-slaughter@ouhsc.edu

CONVERSION OF TWO CONTINUOUS MEASURES USING TRUNCATED REGRESSION: BARTHEL AND FIM SCORES OF STROKE PATIENTS

Yongsung Joo*, University of Florida and V.A. Hospital
Keunbaik Lee, University of Florida
George Casella, University of Florida
Sooyeon Kim V.A. Hospital
Pamela Duncan, V.A. Hospital

A few measures have been used in practice to measure medical condition of stroke patients. Naturally, necessity of studying the relationship between stroke measures comes out when researchers wish to combine the results of papers that used different stroke measures. This paper proposes to use the heteroscedastic truncated regression model for the conversion of two measures. The ECM algorithm is used for estimation.

email: yjoo@biostat.ufl.edu

ANALYZING THE EFFECT OF COVARIATES ON GAP TIMES BETWEEN INTERVAL-CENSORED RECURRENT EVENTS

Dan Sheng*, New York University
Mimi Kim, Albert Einstein College of Medicine

Prospective biomedical studies frequently involve the monitoring of recurrent events. When the occurrences of the event can only be determined through periodic assessments or laboratory tests performed at pre-scheduled visits, the event times can sometimes be interval- censored. In evaluating the effect of covariates on the gap times between interval-censored recurrent events, the complicating issue is that the gap time in between two events may be doubly interval-censored. To address this problem, we propose a method based on a discrete analogue of the proportional hazards model for the marginal distribution of each event. A robust estimator for the variance-covariance matrix is utilized to account for the potential within-subject correlation in the gap times between multiple events. An extensive simulation study indicates that the proposed method yields estimates which are less biased than a midpoint imputation approach. The proposed method is applied to data from the Systemic Lupus Erythematosus National Assessment (SELENA) study.

email: ds744@nyu.edu

---

18. MODELING METHODS IN EPIDEMIOLOGY

MODELING PROSTATE CANCER INCIDENCE: RACIAL DIFFERENCES

Aniko Szabo*, Huntsman Cancer Institute
Alexander D. Tsodikov, University of California, Davis

Prostate cancer is characterized by large racial differences in both incidence and post-diagnosis survival. Known disparities in health care access and utilization complicate uncovering underlying differences in the natural history of the disease. Such differences can be caused by racial variations in risk factors such as diet or genetic predisposition. Understanding and quantifying them is important for the development of health policies. We propose to use modeling techniques to estimate the manner and extent of racial differences in the natural history of prostate cancer. Our model uses the dynamics of prostate cancer incidence after the introduction of prostate- specific antigen (PSA) screening to obtain information about the preclinical stage of the disease.

email: aniko.szabo@hci.utah.edu

A BAYESIAN HIERARCHICAL MODEL FOR ESTIMATION OF DISEASE INCIDENCE USING TWO SURVEILLANCE DATA SETS

Joan Buenconsejo*, Yale University/Food and Drug Administration
Durland Fish, Yale University
Theodore Holford, Yale University
James E. Childs, Yale University/Centers for Disease Control and Prevention

Surveillance data relies on reporting of cases by practicing physicians and it is well recognized that serious underreporting can result. Thus, a model-based approach that allows for the analysis of data from two sources is developed. Such data are comprised of a time series of disease counts, each representing a specific geographical area. A Bayesian hierarchical model is described for estimating the total number of cases with disease in a region. The model simultaneously adjusts for spatial correlation in the incidence, as well as incorporating the capture-recapture methodology to correct for underreporting of cases. This approach explicitly accounts for model uncertainty and can make use of covariates. Inference is carried out using Markov Chain Monte Carlo simulation techniques in a fully Bayesian framework. The methodology is illustrated using data from the Centers for Disease Control and Prevention on Rocky Mountain spotted fever. The information generated by the results not only could provide knowledge of the spatial distribution of infected ticks which depends almost completely upon human case reports, but it could also have significant public health impact that will assist in developing a control program for reducing the morbidity and mortality caused by this preventable and treatable disease.

email: jbsinfuego@yahoo.com

GENETIC MISCLASSIFICATION IN CASE-CONTROL STUDIES

Christine M. Spinka*, University of Missouri–Columbia
Nilanjan Chatterjee, National Cancer Institute
Raymond J. Carroll, Texas A&M University

Epidemiologic studies investigating the interaction between genetics and environmental factors have become increasingly common in recent years. Sometimes data for these studies are collected using a case-control study design. In this setting we utilize a logistic regression model to estimate the probability of disease and include covariates for the effects of genetic factors, environmental factors and their interaction. We develop a method to estimate these parameters in the presence of genetic misclassification, provided that some information about the rate of misclassification can be obtained. We illustrate our method using a simulation study.

email: spinkac@missouri.edu

MODELING HORMONE REPRODUCTIVE DATA USING COMPLEMENTARY APPROACHES

Irina Bondarenko*, University of Michigan
MaryFran Sowers, University of Michigan
Daowen Zhang, North Carolina State University

An association between longitudinally collected continuous variables and symptoms is a focus of many epidemiological studies. Here, we present two complementary approaches to the analyses of complex longitudinal data, relating reproductive hormones and menstrual bleeding status data collected as a part of the Study of Women's Health across the Nation (SWAN). First, we used a linear mixed model approach to evaluate the relationship between annual hormone measures and concurrently-collected self-reported data about changes in a menstrual pattern. Then, to evaluate the consistency of findings from these models, we considered the relationship between hormone data collected daily over one menstrual cycle and annually reported changes in a menstrual pattern. This complementary analysis was conducted within a framework of the semi-parametric stochastic model developed by Zhang and Lin (JASA, 1998). This approach allows for flexible functional dependence of an outcome variable on covariates using nonparametric regression, while accounting for correlation among observations using random effects. In this presentation, we focus on employing different exploratory techniques, such as kernel density estimation and variograms to gain insights about main and random effect structures including the choice of a stochastic process.

email: ibond@umich.edu

THE IMPACT OF THE ANALYTIC APPROACH ON CORONARY CALCIUM AND CARDIOVASCULAR RISK FACTORS

Imke Janssen*, Rush University Medical Center
Zhen Chen, Rush University Medical Center
Lynda H. Powell, Rush University Medical Center

Coronary artery calcification (CAC) may help identify new risk factors for coronary atherosclerosis. The analysis of CAC presents challenges, because the distribution of CAC in asymptomatic populations is highly skewed with many zeros. We review existing approaches for analyzing such data and discuss their advantages and disadvantages. We compare the results on the subclinical disease dataset from the Study of Women's Health Across the Nation (SWAN). A method using survival analysis proposed by Liu (1) yields similar results to binary logistic regression, whereas least squares analysis of ln(CAC+1) does not identify many covariates. Acknowledgements: SWAN is funded by the National Institute on Aging (U01 AG012495, U01 AG012505, U01AG012531, U01 AG012535, U01 A012539, U01 AG012546, U01 AG012553, U01 AG012554), the National Institute of Nursing Research (U01 NR04061) and the NIH Office of Research on Women's Health. (1) Daviglus ML, Liu K, Greenland P, Dyer AR, Garside DB, Manheim L, Lowe LP, Rodin M, Lubitz J, Stamler J (1998). Benefit of a favorable cardiovascular risk-factor profile in middle age with respect to medicare costs. NEJM 339, 1122-1129.

email: Imke_Janssen@rush.edu

DEVELOPING INDICES OF DISEASE SEVERITY IN MORTALITY PREDICTION

Guofen Yan*, Cleveland Clinic Foundation
Tom Greene, Cleveland Clinic Foundation
Gerald Beck, Cleveland Clinic Foundation

In medical studies, disease severity and patient physical status are often included in regression models as covariates in order to better estimate effects of treatment- related factors or to classify homogeneous groups of patients for prediction of outcome. It is often desirable that such tasks be accomplished by a single index. The recently completed Hemodialysis (HEMO) Study, a multi- center NIH supported clinical trial, evaluated the effects of dialysis dose and membrane flux on mortality and morbidity in hemodialysis patients. Mortality was the primary endpoint. Disease severity was evaluated using the summary score of the Index of Coexisting Disease (ICED). The ICED includes 19 individual disease severity scores based on chart review and 11 physical impairment indices. The summary score reduces this information to four levels, ranging from the absence of disease to the severe disease status. We developed several new indices of disease severity using Cox regression. We show that the derived indices improve prediction of mortality and discriminate better between patients with different survival times. Modeling strategies, methodological issues and challenges will be presented.

email: gyan@bio.ri.ccf.org

STATISTICAL MODELING AND INFERENCE FOR HIV/AIDS

Sujay Datta*, Northern Michigan University

Since its first detection in 1981, human acquired immunodeficiency syndrome and its associated pathogen HIV have been the focus of enormous attention, not only because of the worldwide devastation caused by the rapid spread of the virus but also due to the special nature of the infection dynamics that makes it very difficult to find a permanent cure. Compared to other pathogens, an unusually large number of studies have been conducted about the initiation, spread and in-vivo evolution of HIV infection using various mathematical and statistical modeling approaches. The large amount of quantitative data available on various aspects of this infection have motivated these studies. This presentation will be a brief review of some recent developments in mathematical/statistical modeling of, and statistical inference for, this infection. In particular, we talk about pre-ART (ART= anti-retroviral therapy) and post-ART viral dynamics, AIDS clinical trials and so forth. We end with some open problems and future directions of research and provide an extensive list of references.

email: sdatta@euclid.nmu.edu

DETECTING DEPARTURE FROM ADDITIVITY ALONG A FIXED-RATIO RAY WITH A PIECEWISE MODEL FOR DOSE AND INTERACTION THRESHOLDS

Sharon D. Yeatts*, Virginia Commonwealth University
Chris Gennings, Virginia Commonwealth University
Timothy E. O'Brien, Loyola University, Chicago

For mixtures of many chemicals, a ray design based on a relevant, fixed mixing ratio is useful for detecting departure from additivity. Methods for detecting departure involve modeling the response as a function of total dose along the ray. For mixtures with many components, the interaction may be dose-dependent. Therefore, we have developed the use of a three-segment model containing both a dose threshold and an interaction threshold. Prior to the dose threshold, the response is that of background; between the dose threshold and the interaction threshold, an additive relationship exists; the model allows for departure from additivity beyond the interaction threshold. When the interaction threshold is significantly different from the dose threshold, we conclude that there is a region of additivity. The lower confidence bound of the interaction threshold marks the lower bound of the interaction region. The data support of such a complex model may be more efficient when optimal design strategies are used. These designs are dependent on the study objectives, which involve linear or nonlinear hypotheses of model parameters. The methods and designs are illustrated for a mixture of nine chemicals. Supported by T32 ES07334-01A1 (NIEHS, NIH).

email: dziubas@vcu.edu

USING A BAYESIAN HIERARCHICAL MODEL TO ESTIMATE THE RATE OF EMISSION OF GREENHOUSE GASES FROM A FACILITY HOUSING PIGS

Cory R. Heilmann*, Iowa State University
Philip Dixon, Iowa State University

Greenhouse gases, (especially methane, CH4), are a potential byproduct of raising pigs. One approach to estimate rate of emission of CH4 from a facility housing pigs is to introduce a tracer gas not present in nature, SF6, and measure downwind concentrations of CH4 and SF6. We can relate the relative concentrations of these gases to their relative rates of emission. The rate of emission of SF6 is known and we wish to estimate the emission rate of CH4. Since CH4 is present in nature and SF6 is not, we account for the background concentration of only CH4. The statistical problem is to estimate a regression slope with measurement error in the covariates and non-constant variance. Lognormal distributions are placed on the observed concentrations of both gases to remedy non- constant variance. We fit a Bayesian hierarchical model to relate measured concentrations of CH4 and SF6 to the true, unobserved concentration of SF6 based on dilution of both gases downwind from the emission source. We will consider the possibility of different coefficients of variation for the gases as well as different rates of dilution of CH4 and SF6 downwind from the facility. We will present results from one selected day.

email: heilmann@iastate.edu

SPATIAL ESTIMATION OF RISK OF MORTALITY DUE TO AIR POLLUTION

Hae-Ryoung Song*, North Carolina State University
Montserrat Fuentes, North Carolina State University
Sujit Ghosh, North Carolina State University

While many studies have shown evidence of an impact of particulate matter(PM) on human health, there still remains uncertainty about the association between fine particles PM$_{2.5}$ and adverse health effects. In this paper we estimate the impacts of PM$_{2.5}$ and its components on mortality using a Bayesian hierarchical spatio-temporal framework. Based on a generalized Poisson regression model, we adjust for meteorology and several socio-economic confounders. This model is applied to speciated PM$_{2.5}$ and monthly mortality counts over the entire U.S. region for 1999-2000. We obtained a high relative risk of mortality associated to PM$_{2.5}$ in the Eastern and Southern California areas. Regarding the components of PM$_{2.5}$, NO$_3$ and crustal materials have greater health effects in the Western U.S., while SO$_4$ and NH$_4$ have more of an impact in the Eastern U.S. The risk associated to PM$_{2.5}$ is twice what we obtained for PM$_ {10}$.

email: hsong3@unity.ncsu.edu

THRESHOLD REGRESSION AND APPLICATIONS IN ENVIRONMENTAL RESEARCH

Mei-Ling T. Lee*, Harvard University
George A. Whitmore, McGill University

Regression methodology based on observed first hitting times of a threshold for a stochastic process is referred to as threshold regression. For example, the health status of a patient can be assumed to be a latent process. The time to reach the primary endpoint or failure (death, disease onset, etc.) is the time when the latent health status process first crosses a failure threshold level. Threshold regression does not require the proportional hazards assumption and, in many ways, is flexible and realistic for the analysis of time-to-event data. In an application to environmental research, threshold regression was used in a retrospective longitudinal study of more than 50,000 US railroad workers tracked from 1959 to 1996. The initial investigation focused on lung cancer death because of a suspected link to diesel exhaust exposure. Based on the concept that a lung cancer mortality event occurs when the cumulative diesel exposure of a subject first hits a threshold level, a threshold regression model was found to be effective in providing insights into the process of disease progression.

email: meiling@channing.harvard.edu

ESTIMATING CHRONIC EFFECTS OF FINE PARTICLES (PM2.5) ON ADULT MORTALITY AT DIFFERENT SPATIAL AND TEMPORAL SCALES

Sorina E. Eftim*, Johns Hopkins University
Francesca Dominici, Johns Hopkins University
Aidan McDermott, Johns Hopkins University
Scott L. Zeger, Johns Hopkins University
Jonathan M. Samet, Johns Hopkins University

Our scientific objective is to investigate whether long-term (yearly) exposure to ambient fine particles is associated with all-cause mortality in the entire US population of elderly. We have assembled a national data base comprising the Medicare Cohort which includes health death of approximately 40 million participants followed in 2000-2002 matched by county to the National Air Pollution Monitoring Network which includes daily values of PM2.5. We have developed statistical models for estimating associations between county-specific and age-gender-race adjusted all-cause mortality rates versus county-specific long term average PM2.5, separately for each month. Analyses take into account potential confounding by smoking, mobility, and socio-economic status. The methodological approach and findings provide useful information for understanding the effects of ambient particulate matter on human health and for guiding future analyses of particulate data.

email: seftim@jhsph.edu

DATA ANALYSIS AND METHODS FOR THE STUDY OF ENVIRONMENTAL IMPACTS OF ROCKET EMISSIONS

Yi Ye*, University of Missouri–Rolla
Gary L. Gadbury, University of Missouri–Rolla

The potential effects of rocket emissions on climate and stratospheric ozone abundances were studied extensively in recent years. To date, efforts to model the global impacts of rockets have focused on the gas and particulate emissions that deplete stratospheric ozone by enhancing abundances of inorganic chlorine and increasing the reactive surface area of particles. To further study this, CO2 and large particles were measured in the stratospheric plume wakes of Athena II Rocket. One instrument measured CO2 and the other was a laser particle counter (LPC) measuring particulate concentration. The LPC measurements were biased low (termed coincidence loss) at high concentration due to a saturation effect of the instrument. And the placement of the instruments combined with other reasons resulted in the two instruments seeing the same air sample at slightly different times. Statistical methods are developed for aerosol data analysis that involve calibration of coincidence loss, mismatch correction for time differences and detection of saturation point of the LPC instrument. Methodology entails a blending of regression analysis with nonparametic statistical methods such as multi-scale optimization and etc. A simulation study is done to verify the legitimacy of the procedure. Possible applications are on analysis of aerosol particulate data or other complex environmental data.

email: yiy@umr.edu

EMPIRICAL EVALUATION OF SUFFICIENT SIMILARITY IN INFERENCE FOR A MIXTURE OF MANY CHEMICALS USING A FIXED-RATIO RAY DESIGN

LeAnna G. Stork*, Virginia Commonwealth University
Chris Gennings, Virginia Commonwealth University
W. Hans Carter Jr., Virginia Commonwealth University
Robert E. Johnson, Virginia Commonwealth University
Darcy P. Mays, Virginia Commonwealth University

Ray designs based on relevant fixed-ratio rays are useful for testing hypotheses of additivity for large numbers of chemicals. Relevant ratios of chemicals in industrial or environmental processes may not remain constant. While ray designs decrease experimental effort, all possible rays cannot be experimentally evaluated. Interest may be focused on statistical two-sided equivalence tests for additivity or one-sided equivalence tests for lack of synergy along an observed reference ray, and whether the same inference applies to an unobserved candidate ray. Borrowing ideas from the statistical equivalence literature, we have developed methodology that permits us to claim equivalent inference from an observed reference ray to an unobserved relevant candidate ray. Based on this methodology, equivalence margins associated with biologically meaningful deviations are chosen. A confidence region is computed for the reference ray and for the candidate ray based on the variability from the observed data. If both confidence regions are completely contained within the equivalence margins, then the candidate ray may be concluded to be sufficiently similar in inference to the reference ray. The regions of equivalence are depicted graphically for interpretation. This method is illustrated with an example from a mixture of nine chemicals. Supported by T32 ES07334-01A1 (NIEHS, NIH).

email: storklm@vcu.edu

20. ADAPTIVE METHODS AND DESIGNS

ADAPTIVE BAYES DESIGNS FOR DOSE-FINDING PHASE I CLINICAL TRIALS

Yisheng Li, Yuan Ji*, University of Texas M. D. Anderson Cancer Center
Benjamin Bekele, University of Texas M. D. Anderson Cancer Center

We propose a Bayesian adaptive design for dose-finding phase I clinical trials using decision theoretical approaches. We specify a loss function that considers the cost of making wrong decisions for each patient. Specifically, we develop a decision rule that chooses the action that minimizes the Bayesian predicitve loss for the three possible actions that continue the trial, including dose escalation, de-escalation, staying at the same dose level. We also compute the Bayesian posterior loss for the action of trial termination. The trial is terminated only when the loss of trial termination is less than each of the Bayesian predictive losses for the three actions. We derive several theoretical results for the proposed design, and perform extensive simulations to evaluate its operating characteristics. The proposed design is compared with several standard methods and desired results are obtained.

email: yuanji@mdanderson.org

ADAPTIVE DESIGN FOR CENSORED SURVIVAL DATA ADJUSTING FOR COVARIATES

Jie Yang*, University of Florida
Pei-Yun Chen, Merck Research Laboratories
Kaifeng Lu, Merck Research Laboratories

In many clinical trials, it is often of interest to compare time to an event between treatment groups. Most clinical trials of this type are designed with a fixed number of patients or number of events of interest. However, when the observed failure rates at an intermediate stage seem to deviate from the initial assumptions at the planning stage, it is desirable to adjust the sample size in order to have adequate power to detect the smaller yet clinically important treatment difference. We propose an adaptive design of sample size reassessment for censored survival data to take important covariates into consideration. The joint distribution of the score statistics from a proportional hazards model at different stages is derived and its properties are studied. Fisher's variance spending method is applied to construct the test statistic. Simulation studies are carried out to show the operating characteristics of the proposed method under different scenarios and to compare with a fixed sample design and the adaptive design using the weighted log-rank test proposed by Shen and Cai (2003). The method is also illustrated using clinical trial data.

email: jyang2@stat.ufl.edu

ADAPTIVE MODEL-BASED DESIGNS FOR DOSE-FINDING STUDIES

Vladimir Dragalin*, GlaxoSmithKline
Valerii Fedorov, GlaxoSmithKline

We present a class of models that can be used in early phase clinical trials in which patient response is characterized by two dependent binary outcomes, one for efficacy and one for toxicity. We model the distribution of this bivariate binary endpoint using either Gumbel bivariate logistic regression or Cox bivariate binary model. In both cases, the analytic formulae for the Fisher information matrix are obtained, that form the basis for derivation of the locally optimal and adaptive designs. The proposed procedure is based on constrained optimal design theory. Great flexibility can be achieved by using an appropriate penalty function, tailored to the desired goal of the considered clinical trial. Adaptive design has higher efficiency compared with the up-and-down designs. Accounting for toxicity and efficacy response addresses the ethical concern that, as much as possible, subjects be allocated at or near doses that are both safe and efficacious. This design has also a great potential to accelerate the drug development process by combining the traditional Phase I and II in a single trial.

email: Vladimir.2.Dragalin@gsk.com

ADAPTIVE TWO-STAGE DESIGNS IN PHASE II CLINICAL TRIALS

Anindita Banerjee*, North Carolina State University
Anastasios A. Tsiatis, North Carolina State University

Two-stage designs have been widely used in phase II clinical trials. Such designs are desirable because they allow a decision to be made on whether a treatment is effective or not after the accumulation of the data at the end of each stage. Optimal fixed two-stage designs, where the sample size at each stage is fixed in advance, were proposed by Simon (1989) when the primary outcome is a binary response. This paper proposes an adaptive two-stage design which allows the sample size at the second stage to depend on the results at the first stage. Using a Bayesian decision theoretic construct, we derive optimal adaptive two-stage designs; the optimality criterion being minimum expected sample size under the null hypothesis. Comparisons are made between Simon's two-stage fixed design and the new design with respect to this optimality criterion.

email: abanerj2@stat.ncsu.edu

AN ADAPTIVE SINGLE-STEP FDR PROCEDURE WITH APPLICATIONS TO DNA MICROARRAY ANALYSIS

Vishwanath Iyer*, Bristol Myers Squibb
Sanat Sarkar, Temple University

The concept of looking at combined measures of error when testing multiple hypotheses has been increasingly researched by statisticians. Especially in areas such as microarray analysis where the number of hypotheses being tested could be in the thousands, there has been an increasing trend to consider less conservative methods of error control than the ones traditionally used. The false discovery rate is one such error rate, and in this paper, we propose using a method that controls the FDR. The proposed single-step method is in fact an asymptotic approximation to a step-wise procedure that has previously been shown to control the FDR. The method is adaptive in the sense that it estimates the necessary parameters from the data in order to compute a critical value to test the data. It is also shown that this proposed procedure compares favorably with some standard procedures in terms of power and rejection rates. The application of the new procedure to two problems in DNA microarray analysis is also discussed.

email: vishwanath_i@yahoo.com

WHEN ARE ADAPTIVE DESIGNS APPROPRIATE?

Cyrus R. Mehta*, Cytel Software Corporation

Adaptive clinical trials are trials in which interim results, possibly combined with external evidence, are used to modify design parameters for the remainder of the study. The typical trial modification is a re-assessment of the sample size. There is some controversy, however, regarding the appropriateness of the adaptive design because of concerns that more traditional group sequential designs can achieve the same ends with greater efficiency. In this talk we will present some case studies of real situations where the group sequential approach would be impractical to implement whereas the adaptive approach might be helpful in designing a successful trial. Methods for preserving the type-1 error, determining the new sample size and estimating the parameters at the end of an adaptive trial will also be discussed. The presentation will conclude with some comments on the logistical and implementational difficulties that must be resolved in order for adaptive designs to be acceptable in a regulatory environment.

email: mehta@cytel.com

---

21. ANALYSIS OF CORRELATED DATA

THE SIMULTANEOUS ANALYSIS OF MIXED TYPES OF OUTCOMES USING NONLINEAR THRESHOLD MODELS

Todd Coffey*, Virginia Commonwealth University
Chris Gennings, Virginia Commonwealth University

Multiple outcomes are commonly measured on each experimental unit in biomedical studies. The outcomes are typically correlated, and a statistical analysis that incorporates the association may result in improved inference. We propose a methodology to simultaneously analyze binary, count, and continuous outcomes with nonlinear threshold models that incorporates the intra- subject correlation. The methodology uses a quasi- likelihood framework and a working correlation matrix, and is appropriate when the marginal expectation of each outcome is of primary interest and the correlation between endpoints is a nuisance parameter. Parameter estimates are found by solving generalized estimating equations and confidence intervals are based on a generalized score test. Because the derivatives of threshold models do not exist at each point of the parameter space, we outline the necessary modifications that result in asymptotically normal and consistent estimators. Using data from a toxicology experiment, the methodology is illustrated by analyzing five outcomes of mixed type with nonlinear threshold models. Supported by T32 ES07334-01A1 (NIEHS, NIH).

email: coffeyjt2@vcu.edu

GOODNESS-OF-FIT TESTS FOR BINOMIAL GENERALIZED ESTIMATING EQUATIONS (GEE) MODELS: SIMULATION RESULTS

Huiyi Lin*, Tulane University
Leann Myers, Tulane University

Binary outcomes are very common in medical studies. The generalized estimating equations (GEE) methods are often used to analyze correlated binary data. Several goodness- of-fit (GoF) statistics for the GEE methods have been developed recently (Barnhart & Williamson, 1998; Horton et al., 1999; and Pan, 2002). The objective of this study was to compare the existing GEE GoF statistics using simulated data under different conditions. The results from these simulations show that no single GEE GoF statistic performed best under all conditions. Generally, the larger the sample sizes, the more powerful the GEE GoF statistics. The GEE GoF statistics with correctly specified working correlation matrices tend to be robust in terms of Type I error rate and be more powerful. All of the GoF statistics were poor in detecting the omission of the binary time-dependent main variable. Pan's statistics had the best performance for detecting the omission of the interaction for two binary covariates. Barnhart's statistics were the most powerful in detecting the omission of interaction for a time-independent dichotomous variable and a time-dependent continuous variable, the omission of the interaction for a time- independent dichotomous variable and a time-independent continuous variable and the omission of the interaction for two continuous variables.

email: yi0407@yahoo.com

ESTIMATING EQUATION APPROACH FOR TRUNCATED COVARIATES

Gina M. D'Angelo*, University of Pittsburgh Graduate School of Public Health
Lisa Weissfeld, University of Pittsburgh Graduate School of Public Health

Truncated and censored data methodology has been developed for the last 30 years with the focus on the outcome variable. This has lead to the development of models such as the Tobit regression model for censored data in addition to numerous models for censored and truncated data. Another commonly encountered problem is that of truncated covariate data. This type of data is generally observed in the laboratory setting, where the lower limit of detection of an assay is often observed. To address this problem, we propose a method to estimate the coefficients and their standard errors for a regression model with a left truncated covariate using estimating equation techniques. This method will be compared to a standard method of filling in the truncated values with the lower threshold value. The application of this method is illustrated in a sepsis study conducted at the University of Pittsburgh. One aim of this study is to determine the relationship between death status and measures of inflammation such as tumor necrosis factor and interleukin-6.

email: dangelo@upci.pitt.edu

COMPARISON OF WANG-CAREY ESTIMATION VERSUS QUASI-LEAST SQUARES

Wenguang Sun*, University of Pennsylvania School of Medicine
Justine Shults, University of Pennsylvania School of Medicine

In a recent publication (JASA, 2004) Wang and Carey presented a new approach for estimation of the correlation parameters in the framework of generalized estimating equations (GEE). They considered correlated continuous, binary, and Poisson data with a generalized Markov correlation structure. This structure is appropriate for outcomes that stabilize over time and includes the widely used AR(1) and Markov structures as special cases. Wang and Carey (2004) made detailed comparisons with the first stage of quasi-least squares (QLS), a two-stage approach developed in a series of three manuscripts. In this note we complete their assessments by extending them for comparison with the final (bias corrected) version of QLS. We comment on the earlier comparisons, which were overwhelmingly in favor of the Wang-Carey approach. We then show that although the two methods are identical for equally spaced data with an AR(1) structure, they are not equal for unequally spaced data with a Markov structure. Furthermore, we demonstrate that neither (bias corrected) QLS nor the Wang-Carey method is uniformly superior for the scenarios considered by Wang and Carey (2004). We conclude with an application of both methods in an analysis and some general remarks regarding the relative merits of each approach.

email: wsun@cceb.upenn.edu

ADJUSTED QUASI-LEAST SQUARES FOR VALID ANALYSIS OF CORRELATED BINARY DATA

Justine Shults*, University of Pennsylvania School of Medicine
Wenguang Sun, University of Pennsylvania School of Medicine

Recently, concern has been raised regarding violation of bounds for binary data: It is well-known that the correlation among binary outcomes is constrained by the marginal means (Prentice, 1988), yet approaches such as GEE do not check that these conditions are satisfied. In this presentation we first describe when the violation of bounds is likely to occur, especially with regard to sample size, value of the correlation, misspecification of the underlying correlation structure, and violation of an assumption of a constant pattern of association. We consider if a violation of bounds could ever be beneficial and discuss an exploratory approach that could be implemented prior to fitting a patterned correlation matrix. We then propose and demonstrate the benefit of two adjustments to quasi-least squares (QLS) (an approach in the framework of GEE) that can be helpful in overcoming the potential problem of violation of bounds: (1) modifying QLS so that it yields an estimate of correlation that does indeed satisfy the constraints; and (2) allowing the correlation to vary according to subject level covariates. Our methods are developed in the context of a randomized clinical trial, for an outcome variable that we prove will always have an AR(1) correlation structure.

email: jshults@cceb.upenn.edu

ESTIMATION OF CLUSTERED POISSON REGRESSION WITH RANDOM INTERCEPTS

Eugene Demidenko*, Dartmouth Medical School

We compare five methods for parameter estimation of Poisson regression model with cluster-specific intercepts: (1) ordinary (naive) Poisson regression (OP) which ignores intra-cluster correlation, (2) Poisson regression with fixed cluster-specific intercepts (FI), (3) generalized estimating equations (GEE) approach with equi-correlation matrix, (4) exact generalized estimating equations (EGEE) approach with an exact covariance matrix, and (5) maximum likelihood (ML). We prove that these methods produce the same estimates of slope coefficients for balanced data (equal number of observations in each cluster and the same vectors of covariates). All five methods lead to consistent estimates of slopes but have different efficiency for unbalanced data design. The beauty of Poisson regression is that the exact cluster covariance matrix can be derived in closed form without specifying the distribution of the intercepts. Use this matrix in estimating equations constitutes the basis for the EGEE. It is shown that the FI approach can be derived as is a limiting case of maximum likelihood when the cluster variance increases to infinity. In terms of asymptotic efficiency, the methods split into two groups: OP & GEE and EGEE & FI & ML. Thus, contrary to the existing practice, there is no advantage of using GEE because it is substantially outperformed by EGEE or FI.

email: eugened@dartmouth.edu

---

22. LINKAGE ANALYSIS

ON FAMILY-BASED GENETIC ANALYSIS ALLOWING FOR MISSING PARENTAL INFORMATION

Jing Han*, New York University School of Medicine
Yongzhao Shao, New York University School of Medicine

In the collected families in practice, there might be some that the parental information is available and others that the parental information is unavailable. To use incomplete parental information for the family-based genetic analysis, we have proposed a likelihood ratio test by extending the Disequilibrium Maximum-Likelihood- Binomial method of Huang and Jiang (1999, Am. J. Hum. Genet.) and suggested an Expectation-Maximization algorithm for mixture model to estimate missing information on phase and mating type. In this study, we show how all the data can be used jointly in a natural way in one overall likelihood model. The proposed approach is applicable without knowledge of disease penetrance, population allele frequencies, mating types, and the amount of linkage disequilibrium. In addition, the new approach accounts naturally for multiplex affected siblings. The simulation results show that the new approach offers better statistical power than its competitors.

email: jh828@nyu.edu

LINKAGE ANALYSIS OF ORDINAL TRAITS FOR PEDIGREE DATA

Rui Feng*, Yale University
James F. Leckman, Yale University
Heping Zhang,Yale University

Linkage analysis is used routinely to map genes for human diseases and conditions. The existing linkage analysis methods require the corresponding traits to be binary or quantitative. However, many diseases and conditions, such as cancer and mental health conditions, are rated on ordinal scales. The objective of this study was to establish a framework to conduct linkage analysis for ordinal traits. We proposed a latent-variable proportional- odds model that relates inheritance patterns to the distribution of the ordinal trait. We used the likelihood- ratio test for testing evidence of linkage. Through simulation studies, we found that the power of our proposed model is substantially higher than that of the binary-trait- based linkage analysis and that our test statistic is robust with regard to certain parameter misspecifications. Using our proposed method, we performed a genome scan of the hoarding phenotype in a dataset with 53 nuclear families, collected by the Tourette Syndrome Association International Consortium for Genetics. Standard linkage scans were also performed using programs GeneHunter and Allegro and failed to reveal any marker significantly linked to the binary hoarding phenotypes. However, our method identified three markers at 4q34-35 (P=0.0009), 5q35.2-35.3 (P=0.0001), and 17q25 (P=0.0005) that manifest significant allele sharing.

email: rui.feng@yale.edu

JOINT MODELING OF LINKAGE AND ASSOCIATION: IDENTIFYING SNPs RESPONSIBLE FOR A LINKAGE SIGNAL

Mingyao Li*, University of Michigan
Michael Boehnke, University of Michigan
Goncalo R. Abecasis, University of Michigan

An important problem in gene mapping studies is to evaluate whether an observed linkage signal can be explained in part or fully by a SNP that shows evidence of association. We propose a novel approach that quantifies the degree of linkage disequilibrium (LD) between the SNP and the putative disease locus through joint modeling of linkage and association. We describe a simple likelihood for a sample of affected sib pairs with disease penetrances and disease-SNP haplotype frequencies as parameters. We propose two likelihood ratio tests to distinguish the relationship of the SNP and the disease locus. The first test assesses whether the SNP and the disease locus are in linkage equilibrium so that the SNP plays no causal role in the linkage signal. The second test assesses whether the SNP and the disease locus are in complete LD so that the SNP or a marker in complete LD with it may account fully for the linkage signal. Our method also yields a genetic model including parameter estimates for disease-SNP haplotype frequencies and the degree of disease-SNP LD. Our method provides a new tool for detecting both linkage and association and can be extended to study designs that include unaffected individuals.

email: myli@umich.edu

DETECTION OF PLEIOTROPIC GENETIC EFFECTS IN QUANTITATIVE MULTIVARIATE LINKAGE ANALYSIS

Mariza de Andrade*, Mayo Clinic College of Medicine
Curtis Olswold, Mayo Clinic College of Medicine
Stephen T. Turner, Mayo Clinic College of Medicine

Multivariate quantitative linkage analysis provides an approach to identify genes influencing two or more correlated traits. A major advantage over separate univariate analyses is the greater statistical power to identify loci whose effects are too small to be detected in single-trait analyses. Recently we published the first trivariate genome scan for quantitative linkage analysis using blood pressure measures and body mass index data from the Rochester Family Heart Study. Only one region on chromosome 10 showed strong evidence of linkage. Since the multivariate linkage analysis for quantitative traits can be computationally intensive, it is important to develop screening tests to help identifying combinations of traits with shared genes for multivariate analysis by using measures of correlation prior to multivariate linkage analysis. Thus, we propose a test to investigate whether there is a single gene responsible for a particular combination of traits (pleiotropy). This test is simple, fast, and uses the QTL variance component estimates from the univariate quantitative linkage analyses. We applied this test using the univariate linkage analysis results from the blood pressure measures and body mass index data from the Rochester Family Heart Study, and we observed a strong agreement between our test and the trivariate quantitative linkage analysis.

email: mandrade@mayo.edu

MODEL FREE LINKAGE ANALYSIS WHEN THE NUMBER OF ALLELES SHARED IDENTICAL BY DESCENT BETWEEN RELATIVE PAIRS IS MISSING

Tao Wang*, Case Western Reserve University
Robert C. Elston, Case Western Reserve University

Model-free linkage methods are based on the fundamental concept of the number of alleles shared identical by descent (IBD) by relative pairs. Because the markers used in a linkage study are not completely informative, the number of alleles shared IBD can be missing to a different extent for various relative pairs. Schork and Greenwood (2004) have pointed out that the conventional way of imputing the expected value of this number under the null hypothesis of no linkage can introduce a bias toward the null that underestimates this number. In this paper, we show that this imputation bias may inflate the variance of a model-free linkage test statistic, the extent depending on the statistic used. For common complex diseases, the inflated variance for the most often used linkage statistics is limited and therefore unlikely to be one of the major reasons for most failures of model-free linkage studies. We also note that the approach of dropping or down- weighting less informative relative pairs in order to adjust for this bias should be done cautiously as it may result in an invalid statistic. The best way to avoid loss of power is to minimize the ambiguity of allele transmission by using densely spaced markers, which the availability of SNP maps now makes possible.

email: txw54@case.edu

COMBINING EVIDENCE FROM LINKAGE AND ASSOCIATION STUDIES USING DEMPSTER-SHAFER THEORY

Chun Li*, Vanderbilt University
Dan Hahs, Vanderbilt University

In disease gene discovery, investigators often face the challenge of summarizing different sources of information, some confirmatory and some others contradictory. Often, contradictory results are explained using arguments of heterogeneity or sampling variation. If heterogeneity is not an issue and designs are similar across studies, meta-analysis may be used to consolidate results. However, there hasn't been any method on integrating results from studies with sufficiently different designs but the same goal of finding disease-predisposing genes, for example, linkage and association studies conducted on independent data sets from the same population. In this situation, putatively associated genes may or may not be in linked regions, and the levels of association and linkage also vary substantially. Methods are needed to summarize all this information in order to decide which regions or genes to follow up. To address this question, we apply Dempster-Shafer theory (Dempster 1968; Shafer 1976), which has been used in expert systems to combine degrees of belief derived from independent sources of evidence. Using posterior probabilities, we first define degrees of belief on candidate regions based on linkage analysis and on candidate genes based on association analysis. Then, we apply Dempster-Shafer theory to combine these degrees of belief so that investigators have an overall picture of current findings. We will describe the method and apply it to real data.

email: chun.li@vanderbilt.edu

A LOGISTIC REGRESSION MIXTURE MODEL FOR INTERVAL MAPPING OF GENETIC TRAIT LOCI AFFECTING BINARY PHENOTYPES

Weiping Deng*, George Washington University
Hanfeng Chen, Bowling Green State University
Zhaohai Li, George Washington University

We propose a method by combining interval mapping with logistic regression for mapping genetic trait loci affecting binary disease traits. We call the trait loci for the binary phenotypes as binary trait loci (BTL). To achieve the model identifiability, we assume that the presence or absence of a disease is influenced by not only the trait locus genotypes but also some environmental condition which is incorporated in our model as a covariate. This covariate helps identify the logistic regression mixture model by newly defined parameters. Because of the irregularity of mixture models, the usual chi-square approximation to likelihood ratio test (LRT) statistic is not applicable. Instead, under our assumption for the covariate, its null limiting distribution is a supremum of chi-square processes which is used to determine the critical value of the LRT and to test the hypothesis for detection of BTL. The method is illustrated by simulated data from a backcross design to test the effect of BTL.

email: wpd@gwu.edu

23. STATISTICAL ISSUES AND NOVEL METHODS IN VACCINE CLINICAL TRIALS

EVALUATING HIV VACCINE: SELECTING ENDPOINTS AND VALIDATING SURROGATES

Thomas Fleming*, University of Washington

In this talk, we will discuss the criteria for valid endpoints in clinical trials and illustrate them using examples in the HIV vaccine trials. We point out that a correlate of a clinical endpoint may not be a surrogate endpoint. We will further discuss the validation of surrogate endpoints and the associated controversial issues with the accelerated approval.

email: fleming@seattlebiostat.com

CAUSAL VACCINE EFFECTS ON BINARY POST-INFECTION OUTCOMES

Michael G.Hudgens, University of North Carolina at Chapel Hill
M. Elizabeth Halloran*, Emory University

Evaluation of many effects of prophylactic vaccines on outcomes such as severe disease, death, or transmission to others, condition on being infected. Conditioning on an event that occurs posttreatment, in our case infection subsequent to assignment to vaccine or control, could result in selection bias, because the people who become infected in the vaccinated group might not be comparable to those who become infected in the unvaccinated group. In this talk, we consider identifiability and estimation of causal effects of vaccination on binary post-infection outcomes such as transmission to others, severe disease, and death. We use the Frangakis and Rubin (2002) approach to define causal effects within principal strata of individuals who have the same joint potential infection values under vaccine and control. We develop a likelihood model to define and to estimate the causal estimands. We derive closed forms for the MLEs of the post-infection causal estimates under the extreme selection models and present three methods for sensitivity analyses under varying assumptions of the selection bias. We analyze data from studies of rotavirus and pertussis vaccines.

email: mehallo@sph.emory.edu

DEMONSTRATING THAT AN HIV VACCINE LOWERS THE RISK AND/OR SEVERITY OF HIV INFECTION

Devan V. Mehrotra*, Merck Research Laboratories
Xiaoming Li, Merck Research Laboratories
Peter B. Gilbert, Fred Hutchinson Cancer Research Center

Consider a placebo-controlled clinical trial to assess whether an experimental HIV vaccine reduces the risk of HIV infection and/or the "severity" of HIV infection (viral load set-point) in those who become infected despite vaccination. The null hypothesis is that the vaccine does not reduce the incidence or severity of HIV infection. A method for testing this hypothesis was proposed by Chang et al (1994). In their method, for each randomized group, a burden-of-illness (BOI) per randomized subject is calculated by dividing the sum of the severity scores of all infected subjects by the number randomized. A test statistic is then used to evaluate the statistical significance of the between-group difference in BOI. In this talk, we compare the BOI method with two alternative methods (Simes' and Fisher's) that combine p-values from comparative tests for the incidence and (post-infection) severity components. The latter methods are shown to be generally more powerful than the former. This result holds even after incorporating a "penalty" for potential selection bias when comparing the between-group severity scores of subjects that become HIV infected. The penalty is derived using the selection bias model proposed by Gilbert et al (2003), based on Rubin's causal inference framework.

email: devan_mehrotra@merck.com

---

24. STATISTICAL ANALYSIS OF WILDFIRE DATA

ESTIMATING WILDFIRE MANAGEMENT EFFECTIVENESS

David T. Butry*, USDA Forest Service
Marcia L. Gumpertz, North Carolina State University
Marc G. Genton, Texas A & M University

We regress wildfire size in the St. Johns River Water Management District in Florida on climate and weather variables, cause of fire, natural and socioeconomic landscape characteristics, and wildland management, taking spatial and temporal relationships into account. This analysis provides evidence that large (>1,000 acres) wildfires are related to a different suite of factors than small wildfires, perhaps indicating the need for different mitigation strategies. For instance, we find that fuels management, i.e. preventive burning, is associated with reduced wildfire size, but only for small wildfires. This result is intriguing, however it is difficult to quantify the effect of fuels management on fire size or attribute causality to it. In the social sciences, matching on propensity scores has been found useful for evaluating program success from observational data where randomized assignment of treatments is impractical. In propensity score matching, units with similar characteristics are matched, then the responses of those who participated in the program are compared with those who did not. We explore the use of propensity score matching to estimate the effect of preventive burning on the area averted from wildfire.

email: gumpertz@ncsu.edu

TOWARDS IMPROVED PREDICTION OF WILDFIRE RISK

Frederic P. Schoenberg*, University of California, Los Angeles
Maria Chang, Haiyong Xu, University of California, Los Angeles
Jamie Pompa, University of California, Los Angeles
James Woods, California State University, Long Beach
Roger D. Peng, Johns Hopkins University

Wildfires pose an extremely serious threat to Southern California each year, damaging vast areas of public and private land and often resulting in massive losses of property and threatening human lives. In 2003, between October 24 and November 2 alone, wildfires burned nearly 700,000 acres, causing the destruction of more than 3,300 homes and the deaths of at least 20 people. The Burning Index (BI) produced by the U.S. National Fire Danger Rating System is a complex amalgam of meteorological variables designed to forecast wildfire hazard, but empirically is a rather poor predictor of wildfire activity in Los Angeles County. An alternative is to make direct use of meteorological and historical wildfire information, including temperature, precipitation, fuel age, and wind. We will focus on the efficiency of these variables in predicting wildfires, and will explore the use of spatial-temporal point process models that incorporate this information and their predictive capabilities.

email: frederic@stat.ucla.edu

FIRE REGIMES: CONTROLS AT DIFFERENT SCALES OF SPACE AND TIME

Max A. Moritz*, Ecosystem Sciences, ESPM, University of California, Berkeley

Fire is one of the most fundamental forces affecting patterns and processes in terrestrial ecosystems, and there are several competing factors that may drive the “fire regime” of a region. The importance or strength of these factors can vary over space and time, making it difficult (and sometimes inappropriate) to generalize about patterns of fire behavior and their controls. Further complicating matters is the fact that there are several different parameters that are typically measured when describing the fire regime of a region. In my talk I will give a general overview of how fire regimes are characterized and quantified, in addition to examples of how simplistic models can lead to misunderstandings and endless debate. I will also illustrate how identification of controls on a fire regime can vary, depending on the scale of analysis, using case studies from my own work.

email: mmoritz@nature.berkeley.edu

MODELING SIZE OF LARGE CATASTROPHIC WILDFIRES USING SKEW-ELLIPTICAL DISTRIBUTIONS

Marc G. Genton*, Texas A&M University

We consider a data set provided by the USDA-Forest Service about the size of 111 large (>1,000 acres) catastrophic wildfires located in the St Johns River Water Management District of northeastern Florida. A preliminary empirical analysis indicates that a wildfire becomes large only when certain environmental characteristics are met. This suggests that an underlying latent process, possibly spatially distributed, determines the development of large wildfires. That is, the selection of a fire into our sample induces skewness. We develop selection models for spatial data based on multivariate skew-elliptical distributions. These models provide the basis of a new skew kriging procedure that allows to incorporate skewness in spatial interpolation, and reduces to the traditional kriging equations when there is no skewness, i.e. no latent process. This is joint work with A. Dominguez-Molina and G.Gonzalez-Farias.

email: genton@stat.tamu.edu

25. BAYESIAN STATISTICAL MODELING OF MASS SPECTROMETRY PROTEOMIC DATA

SOURCES OF VARIABILITY IN MALDI-TOF MS PROTEIN PROFILING

Dean Billheimer*, Vanderbilt University

Matrix-assisted laser desorption-ionization, time-of-flight (MALDI-TOF) mass spectrometry (MS) is a leading technology in proteomics. This technology allows direct measurement of the protein 'signature' of tissue, blood, or other biological specimens, and holds tremendous potential for disease screening, diagnosis and treatment. Despite recent technical advances in signal generation, several factors intrinsic to the measurement process contribute to variation in the observed spectrum. These characteristics include variation in mass/charge assignment, non-zero baseline, multiplicative intensity scaling, and mass/charge dependent detector efficiency. I provide a summary of the MALDI-TOF MS technology, and describe statistical characteristics of these data. I focus on the estimation and removal of nuisance parameters (e.g., baseline, scaling). These data 'pre-processing' steps are frequently based on ad hoc procedures, and implemented by convenience rather than by justifiable estimation methods. I submit that such data transformations are the most important steps in analysis of MALDI-TOF MS data. Further, nuisance parameter accommodation should be based on defensible statistical principles.

email: dean.billheimer@vanderbilt.edu

A BAYESIAN MIXTURE MODEL FOR PROTEIN BIOMARKER DISCOVERY

Kim-Anh Do*, University of Texas M. D. Anderson Cancer Center
Peter Mueller, University of Texas M. D. Anderson Cancer Center
Raj Bandyopadhya, Rice University

Early detection is critical in disease control and prevention. Biomarkers provide valuable information about the status of a cell at any given time point. Biomarker research has benefited from recent advances in technologies such as proteomics. Motivated by specific problems involving proteomic profiles generated from mass spectrometry techniques, we propose model-based inference based on mixtures of beta distributions for real-time discrimination in the context of protein biomarker discovery. Key to the proposed inference is a perspective of recognizing the relevant sampling distribution as a density estimation problem. The use of a density estimation likelihood is different from the more common approach of considering the problem as one of smoothing the noisy observed ``spectrum'' by essentially nonlinear regression. We assume for each sample a different mixture of Beta random probability model. The mixture models are linked at the level of a hierarchical prior. The prior includes positive prior probability for mixture terms being common across biologic conditions or specific to only one condition. Posterior inference formalizes the desired inference on proteins with a certain mass/charge ratio being differentially expressed across the experimental conditions of interest.

email: kim@mdanderson.org

BAYESIAN NONPARAMETRIC MODELS FOR PROTEOMIC EXPRESSION

Merlise A. Clyde*, Duke University
Leanna House, Duke University
Robert Wolpert, Duke University

Motivated by the science of expression proteomics, we develop a Bayesian nonparametric approach for modeling mass spectrometry - Time-of-Flight (MS-TOF) data from multiple subjects, where data are individual curves or spectra. Critical steps in data analysis include calibration of protein abundance across individuals, peak or protein detection from noisy data, and identification of peaks or regions of the spectra with differential expression for the goal of classification of subjects. We develop a prior model on functions for protein abundances in individual spectrograms using a Levy random field model that allows an unknown number of proteins. The posterior distribution for the number, location and abundance of proteins in each individual spectra is obtained via a reversible jump Markov chain Monte Carlo algorithm. We build a hierarchical model to identify differentially expressed proteins between diseased and non-diseased groups, and ultimately, make probabilistic statements concerning the predictions of patient disease status. The nonparametric models described above are used to address several statistical issues simultaneously, enabling us to incorporate realistic measures of model uncertainty in the classification of disease status and in reported probabilities of differential protein expression.

email: clyde@stat.duke.edu

BAYESIAN MODELING AND INFERENCE FOR MASS SPECTROMETRY DATA USING FUNCTIONAL MIXED MODELS

Jeffrey S. Morris*, University of Texas M. D. Anderson Cancer Center
Kevin R. Coombes, University of Texas M. D. Anderson Cancer Center
Keith A. Baggerly, University of Texas M. D. Anderson Cancer Center
Philip J. Brown, University of Kent, Canterbury

In this talk, we demonstrate how to analyze mass spectrometry data using the wavelet-based functional mixed model approach of Morris and Carroll (2004), which is a generalization of the linear mixed model to functional data. This approach is very general, accommodating a wide class of experimental designs and allowing one to identify protein peaks related to various outcomes of interest, including dichotomous outcomes, categorical outcomes, continuous outcomes, and any interactions among factors. These factors can be conditions of interest (e.g. cancer/normal) or experimental factors for which we wish to account (blocking factors). Random effects make it possible to model correlation between spectra from the same individual or block. The MCMC output can be used to detect peaks, find which ones are related to factors of interest, and to classify future samples based on predictive probabilities. This method is applied to two MALDI data sets from experiments run at MD Anderson, one a clinical study whose goal is diagnosis of pancreatic cancer from blood serum, and the other an animal study studying the serum proteome of mice injected with one of three cell lines in one of four organs. This methodology appears promising for the analysis of mass spectrometry data.

email: jeffmo@odin.mdacc.tmc.edu

---

26. NEW APPROACHES: TO STATISTICAL ACCESS TO DATA IN A CONFIDENTIAL WORLD

REGRESSION ON DISTRIBUTED DATABASES VIA SECURE MULTI-PARTY COMPUTATION

Alan F. Karr*, National Institute of Statistical Sciences
Xiaodong Lin, National Institute of Statistical Sciences
Ashish P. Sanil, National Institute of Statistical Sciences
Jerome P. Reiter, Duke University

We present a method for performing linear regression on the union of distributed databases that does not entail constructing an integrated database, and therefore preserves confidentiality of the individual databases. The method can be used by statistical agencies to share information from their individual databases, or to make such information available to others.

email: karr@niss.org

BOUNDS FOR CELL ENTRIES IN MULTI-WAY TABLES GIVEN COMBINATIONS OF MARGINALS AND CONDITIONALS

Aleksandra B. Slavkovic*, Pennsylvania State University

Statistical disclosure limitation applies statistical tools to the problem of limiting releases of sensitive information about individuals and groups that are part of statistical databases while allowing for proper statistical inference. A way to achieve this in k-way contingency tables is through partial data releases of sets of marginals and conditionals. We extend recent results on upper and lower bounds for the entries in contingency tables given arbitrary collections of marginals and conditionals, and we discuss some implications of these results for statistical disclosure limitation.

email: abs12@psu.edu

ALGEBRAIC GEOMETRY TOOLS FOR STATISTICAL DISCLOSURE LIMITATION AND STATISTICAL ESTIMATION IN CONTINGENCY TABLES

Alessandro Rinaldo, Carnegie Mellon University
Stephen E. Fienberg*, Carnegie Mellon University
Aleksandra B. Slavkovic, Pennsylvania State University
Seth Sullivant, University of California, Berkeley

Linear and integer programming and network analysis have been among the tools from operations research and mathematics that have been used in the implementation of methods for disclosure limitation. Recently, ideas from computational algebra and polyhedral geometry have been used to derive new procedures for disclosure limitation risk assessment and and analyze maximum likelihood estimation in contingency tables. Here we explore the links between these areas. For example, we explain how the release of margins associated with sparse contingency tables allows for a precise identification of some of the zero entries when the MLEs for corresponding log-linear models do not exist. These zero values further constrain the bounds for other cell entries. We also describe some of the relevant computational resources we have drawn upon and their uses.

email: fienberg@stat.cmu.edu

NON-NORMAL RANDOM EFFECTS IN GENERALIZED LINEAR MIXED MODELS

Alan Agresti*, University of Florida

In the past ten years, much attention has focused on the class of generalized linear models that allow random effects, the 'generalized linear mixed model' (GLMM). It is common to assume that random effects follow a normal distribution, but some literature has considered a nonparametric approach. We describe this approach and give a logistic regression example in which a binary random effects distribution is natural. We also discuss results of a study about possible efficiency loss in assuming normal random effects when the true random effects distribution is nonnormal.

email: aa@stat.ufl.eduu

PARTIALLY OBSERVED INFORMATION AND INFERENCE ABOUT NON-GAUSSIAN MIXED LINEAR MODELS

Jiming Jiang*, University of California, Davis

In mixed linear models with nonnormal data, the Gaussian Fisher-information matrix is called quasi information matrix (QUIM). The QUIM plays an important role in evaluating the asymptotic covariance matrix of the estimators of the model parameters. Traditionally, there are two methods of estimating the information matrix: the estimated information and the observed one. Because the analytic form of the QUIM involves parameters other than the variance components, for example, the third and fourth moments of the random effects, the estimated QUIM is not available. On the other hand, because of the dependence and nonnormality of the data, the observed QUIM is inconsistent. We propose an estimator of the QUIM consisting partially of an observed form and partially of an estimated one. We show that this estimator is consistent and computationally very easy to operate. The method is used to derive large sample tests of statistical hypotheses that involve the variance components in a non- Gaussian mixed linear model. Finite sample performance of the test is studied by simulations and compared with the delete-group jackknife method that applies to a special case of non-Gaussian mixed linear models.

email: jiang@wald.ucdavis.edu

A SEMIPARAMETRIC LIKELIHOOD APPROACH TO GENERALIZED LINEAR MODELS WITH COVARIATES AS RANDOM EFFECTS FOR LONGITUDINAL DATA

Erning Li*, Texas A&M University
Daowen Zhang, North Carolina State University
Marie Davidian, North Carolina State University

A relevant framework to joint model a primary outcome and longitudinal profiles of a continuous response is to assume the longitudinal data follow a mixed model and the primary outcome follows a generalized linear model depending on the underlying random effects for longitudinal data and other covariates. Interest may focus on the association between the primary endpoint and features of longitudinal profiles; the underlying random effects distribution may also be of interest. Normality of the random effects distribution is a routine assumption made by most existing methods for fitting the joint model, but such a restrictive assumption may be unrealistic. We relax this assumption and propose a likelihood-based semiparametric approach which requires only that the random effects have a smooth density. We approximate the random effects distribution by the seminonparametric (SNP) density representation of Gallant and Nychka (1987 Econometrica 55:363-390). EM algorithm is used for implementation. The proposed method yields valid inference and shows potential efficiency gains over competing methods under various random effects distributions. Moreover, it provides reliable estimator for the underlying distribution of random effects. We illustrate the approach by application to data from a study of osteopenia in peri-menopausal women and via simulation.

email: eli@stat.tamu.edu

MAXIMUM LIKELIHOOD INFERENCE IN NON-NORMAL RANDOM EFFECTS MODELS

Peter X. K. Song*, University of Waterloo
Peng Zhang, University of Waterloo
Annie P. Qu, Oregon State University

We present a maximum likelihood inference in linear mixed models when random effects may follow a non-normal distribution. Our inference procedure is built on a family of parametric distributions including the normal as a special case. This thus provides an extension of the classical normal random effects model and allows both heavy tailed and skewed random effects distributions. We develop a general fixed point algorithm for the estimation of both fixed effects and variance component parameters. We also propose a lack-of-fit test to assess the departure of the normality from a given non-normal random effects distribution. To illustrate the proposed methods, we investigate two distribution families for random effects: the multivariate t-distribution and the semi-nonparametric (SNP) distribution families. We provide both simulation studies and data analysis examples, and compare our approach to the existing EM algorithm.

email: song@uwaterloo.ca

ROBUST TREND TESTS FOR GENETIC ASSOCIATION USING MATCHED CASE-CONTROL DESIGN

Gang Zheng, National Heart, Lung, and Blood Institute
Xin Tian*, National Heart, Lung, and Blood Institute

We derive a trend test for genetic association using a matched case-control design which allows for a variable number of controls per case. This trend test depends on the underlying genetic model. Since for many complex diseases the mode of inheritance is unknown, two robust trend tests without specifying genetic models are studied. Simulation is conducted to compare the performance of the trend tests and the robust trend tests under various genetic models. The robust trend tests are applied to detect candidate-gene association using an example from a recent case-control etiologic study of sarcoidosis.

email: tianx@NHLBI.nih.gov

EDUCATION DELAYS ACCELERATED DECLINE IN MEMORY IN PERSONS WHO DEVELOP ALZHEIMER'S DISEASE

Charles B. Hall*, Albert Einstein College of Medicine of Yeshiva University
Carol A. Derby, Albert Einstein College of Medicine of Yeshiva University
Mindy J. Katz, Albert Einstein College of Medicine of Yeshiva University
Aaron J. LeValley, Albert Einstein College of Medicine of Yeshiva University
Joe Verghese, Albert Einstein College of Medicine of Yeshiva University
Herman Buschke, Albert Einstein College of Medicine of Yeshiva University
Richard B. Lipton, Albert Einstein College of Medicine of Yeshiva University

Persons who develop Alzheimer's disease experience an accelerated rate of decline in cogntive ability, in particular memory. Education is believed to be protective against the development of dementia, but the mechanism for this is not well understood. We studied the natural history of memory in 117 participants in the Bronx Aging Study through the use of linear spline models in which the change point is estimated from the data and is allowed to depend on education. We extended the profile likelihood method (Hall et al 2003) to estimate the parameters describing the change point. Results indicated that each additional year of formal education delayed the time of accelerated decline on the Buschke Selective Reminding Test by 0.2 years (95% CI 0.04-0.48 years), but the rate of decline after the change point increased by 0.1 point per year (95% CI 0.02-0.18 points per year) for each year of additional formal education. These findings suggest that education may alter the risk of dementia by delaying the onset of accelerated memory decline during the preclinical course.

email: chall@aecom.yu.edu

POPULATION LAB: THE CREATION OF VIRTUAL POPULATIONS IN GENETIC EPIDEMIOLOGY RESEARCH

Monica Nichifor*, Karolinska Institutet, Stockholm, Sweden
Marie Reilly, Karolinska Institutet, Stockholm, Sweden

To estimate the familial contribution to the risk of familial aggregated diseases valuable information is provided by considering the number of affected relatives, their degree of relationship and age at diagnosis. If such information is recorded in population-based registers, these offer a very efficient means of immediate electronic follow up of study cohorts or identifying cases. However, little work has been done on estimating how the validity of results based on such registers is affected by truncation and incompleteness of family relationships. We present a method and software for simulating populations of related individuals evolving over calendar time. We store this population in a pedigree file. The software package, written for the R environment, is called Population Lab. We demonstrate with a simulation of the Swedish female population for the calendar period 1955 - 2002, using breast cancer incidence as the disease of interest. Standard demographic features agree well with the real population and disease incidence parameters are recovered by appropriate statistical analysis. Population Lab will be further developed as a tool for investigating the impact of missing family members on epidemiological analyses and for developing and testing statistical methods that accommodate incomplete family data.

email: monica.nichifor@meb.ki.se

ATTRIBUTABLE RISK ESTIMATION IN LONGITUDINAL STUDIES WITH CENSORING

Cynthia S. Crowson*, Mayo Clinic
Terry M. Therneau, Mayo Clinic
Sherine E. Gabriel, Mayo Clinic
William M. O'Fallon, Mayo Clinic

Population attributable risk (AR; aka etiologic fraction) is defined as the proportion of a disease that could be prevented by elimination of a risk factor from the population. Conceptually, AR combines the prevalence of a risk factor with the prognostic effect of the risk factor. An epidemiologic or public health measure, AR is typically estimated from case-control or cross-sectional studies. However, it is relevant and can be estimated in a longitudinal study if two issues are properly addressed. First, the prevalence of the risk factor in a longitudinal study cohort is changing over time due to new exposures and to deaths. Second, when subjects are followed long enough longitudinally, some diseases may not be preventable, but may be delayed. These issues can be addressed in Cox proportional hazards models by incorporating time-dependent covariates for the risk factor of interest and adjusting for the competing risk of death. Roughly, AR as a function of time is estimated by comparing the cumulative disease incidence estimated from such a Cox model including the risk factor of interest with the estimate assuming no one in the cohort has the given risk factor.

email: crowson@mayo.edu

HIERARCHICAL MODELS FOR COMBINING ECOLOGICAL AND CASE-CONTROL DATA

Sebastien J. Haneuse*, Vanderbilt University;
Jonathan C. Wakefield, University of Washington

Ecological studies suffer from a range of potential biases which give rise to severe difficulties in the interpretation of their results. The fundamental problem is due to the inability of ecological data alone to characterise within-group exposure distributions. We propose a hybrid study design which combined ecological data with individual-level case-control data. The latter provide within-group information, and hence the basis for the control of bias, while the former may provide efficiency gains. Here, we outline the proposed study design, together with details regarding the corresponding hybrid likelihood. One feature of the approach is that it is based on exact distributions, and is therefore computationally expensive. Frequentist estimation and inference requires careful monitoring of the maximisation process. Adopting the Bayesian statistical paradigm, with the use of Markov Chain Monte Carlo algorithms, leads to considerable simplifications. Specifically, we find the use of auxiliary variable schemes to be of particular use. We illustrate the performance of the hybrid approach, together with various computational issues, using a hierarchical model which incorporates spatial dependence.

email: sebastien.haneuse@vanderbilt.edu

CAUSAL INFERENCE FOR MORBIDITY OUTCOMES IN THE PRESENCE OF DEATH

Brian L. Egleston*, Johns Hopkins University
Daniel O. Scharfstein, Johns Hopkins University
Ellen E. Freeman, Johns Hopkins University
Sheila K. West, Johns Hopkins University

Evaluation of the causal effect of an exposure on a morbidity outcome is often complicated by the presence of death as a competing risk. In this setting, the causal effect is only well-defined for the principal stratum of subjects who would live whatever be the exposure. Motivated by aging researchers interested in understanding the causal effect of vision loss on emotional distress in a population with a high mortality rate, we introduce a set of scientifically driven assumptions to identify the causal effect among those who would live both with and without vision loss. To evaluate the robustness of our analysis to nonidentifiable assumptions, we propose a method for performing a sensitivity analysis. We apply our method using the first three rounds of survey data from the Salisbury Eye Evaluation, a population-based cohort study of older adults.

email: beglesto@jhsph.edu

29. CLINICAL TRIALS I

A SEQUENTIAL PROCEDURE FOR MONITORING CLINICAL TRIALS AGAINST HISTORICAL CONTROLS

Xiaoping Xiong*, St. Jude Children's Research Hospital
Ming Tan, University of Maryland Greenebaum Cancer Center
James Boyett, St. Jude Children's Research Hospital

We propose a sequential procedure for monitoring clinical trials against historical controls which is sometimes the only alternative design for a phase III trial design that is intended by not feasible. When there is a strong ethical concern against the patients to be randomized to the existing treatment, or when the enrollment is too limited to randomize subjects into experimental and control groups, one may monitor the trial sequentially against historical controls if the historical data with required qualities and sample size are available to form a reference scheme for the trial. This type of clinical trials brought forth a statistical problem for comparing two population means in a way that data from one population is sequentially collected and compared with all data from the other population at each interim look. We develop the sequential procedure based on the sequential conditional probability ratio test (SCPRT) by which the conclusion of the sequential test would be the same as that by a nonsequential test based on all data at the end of trial. We develop the sequential procedure by generalizing SCPRT for Brownian motion with dependent increments that emulates the test statistic comparing interim data in the current study sequentially with all data in the historical study.

email: xiaoping.xiong@stjude.org

STATISTICAL APPROACHES FOR EVALUATING NON-INFERIORITY TRIALS WHEN THE NON-INFERIORITY MARGIN DEPENDS ON THE CONTROL EVENT RATE

Mimi Y. Kim*, Albert Einstein College of Medicine
Xiaonan Xue, Albert Einstein College of Medicine

In non-inferiority trials where the outcome is binary, the margin of non-inferiority is often defined as a fixed 'delta', the largest clinically acceptable difference in event rates. This margin is usually determined after taking into consideration the expected event rate in the control group. The problem with the fixed delta approach is that there can be considerable uncertainty surrounding the control event rate so that the defined delta may not be appropriate if the true control rate is much larger or smaller than the assumed rate. An alternative approach is to allow delta to vary according to the true control rate. The appropriate statistical approach for evaluating non-inferiority in this situation is not readily apparent, however, especially when the dependence between delta and the control rate may be non- linear. This paper considers and compares three approaches for evaluating non-inferiority with a variable non- inferiority margin: an observed event rate approach, a Bayesian approach, and a likelihood ratio based test. The methods are illustrated with a non-inferiority trial involving subjects with systemic lupus erythematosus.

email: mikim@aecom.yu.edu

PREDICTING EVENT TIMES IN CLINICAL TRIALS IN THE ABSENCE OF TREATMENT ARM INFORMATION

Mark Donovan*, University of Pennsylvania
Michael R. Elliott, University of Pennsylvania
Daniel F. Heitjan, University of Pennsylvania

Because power is primarily determined by the number of events in event-based clinical trials, the timing for interim or final analyses of data is typically determined based on the accrual of events during the course of the study. It is of interest to predict early and accurately the time of a landmark event (interim or terminating event), for timing of analysis. Existing Bayesian methods may be used to predict the date of the landmark event, based on current enrollment, event, and loss to follow-up if treatment codes are known. This work extends these methods to the case where the treatment codes are masked using a parametric mixture model with a known mixture proportion. The mixture model approach is compared with posterior simulation methods assuming a single population. Comparison of the mixture model with the single population approach shows that with few events, these approaches produce substantially different results, and that these results converge as the prediction time is closer to the landmark event. Simulations show that the mixture model with use of appropriate priors tends to have better coverage probabilities for the prediction interval than the non-mixture models if there is a difference in event rates by treatment arm.

email: jdonovan@cceb.upenn.edu

SIMPLE CONFIDENCE BOUNDS AT EFFECTIVE DOSES IN DOSE FINDING STUDIES

Yi-Hsuan Tu*, Columbia University
Ying-Kuen Cheung, Columbia University

We consider dose finding studies in which several doses are compared against a control group in terms of some efficacy measurements. The objectives of such studies are usually two-fold: (1) to identify the minimum dose that is effective when compared to the control; and (2) to obtain unbiased estimation of efficacy at the minimum effective dose relative to that at the control. While there is a rich statistical literature of multiple comparisons in the many-to-one situations, existing methods do not accommodate both objectives. In this paper, we propose a simple α-splitting approach to construct confidence bounds at effective doses. We present a specific procedure that extends a method (called DR method) discussed in Hsu and Berger (1999). Through simulations and examinations of several data sets, we find that our method gives a more precise confidence bound at the effective doses than the DR method, while maintaining similar power in terms of the selection of effective doses. Our proposed approach is versatile in that it can be applied to extend other stepwise methods that are designed to identify the minimum effective dose. We focus on the situations where dose- response is monotone in the article, but also discuss on how the α-splitting approach may be applied for situations where monotonicity does not hold.

email: yt2011@columbia.edu

EQUIVALENCE ASSESSMENT ON MULTIPLE PROPORTION OUTCOMES

Lan Kong*, University of Pittsburgh
Robert C. Kohberger, University of North Carolina at Chapel Hill
Gary G. Koch, University of North Carolina at Chapel Hill

In clinical equivalence trials, assessment of equivalence of treatments may involve multiple proportion outcomes. For example, vaccine trials usually compare between two vaccines on the percent of subjects whose immune responses are above a certain level that is putatively defined as the `protective' level. We consider a scenario where two treatments have to show equivalence uniformly on each component of the proportion vector. We conduct a simulation study to show how the correlations among the multiple outcomes affect the Type I error and power. We also describe how to compute the power for testing equivalence on multiple responding rates that result from multivariate normal responses based on a prespecified criterion.

email: lkong@pitt.edu

INCORPORATING INTERIM ANALYSES AND/OR HISTORICAL CONTROLS INTO FLEXIBLE SCREENING TRIALS

Daniel J. Sargent*, Mayo Clinic, Rochester
Susan Geyer, Mayo Clinic, Rochester
Haolan Lu, Mayo Clinic, Rochester

The overall goal in evaluating various experimental regimens in the clinical trial setting is to identify a regimen that is as good or better than the current standard of care. Screening trials, in the form of randomized Phase II trials, have been proposed and used to identify which of two or more experimental regimens should be explored further in the Phase III setting. Sargent and Goldberg (2001, Stat in Med) (SG) proposed one such design that allows simultaneous screening of multiple agents. Decisions in this design are based on a flexible decision rule that allows the incorporation of additional factors (such as toxicity or convenience) when the observed difference in the primary endpoint is modest. Here, we propose two extensions to the SG design: 1) the incorporation of a formal interim analysis, and/or 2) the addition of hypothesis testing against historical data (i.e. seeking a minimal level of efficacy). We feel that these additions increase the clinical utility of the SG design, by respectively providing the ability to terminate the trial early if one regimen is clearly more promising than the other(s), and by ensuring that there is a sufficient level of efficacy to warrant a comparison with the standard of care.

email: sargent.daniel@mayo.edu

QUANTIFYING PLACEBO EFFECT IN DISCONTINUATION TRIALS USING FUNCTIONAL DATA

Eva Petkova*, Columbia University
Thaddeus Tarpey, Wright State University
Todd R. Ogden, Columbia University

An important problem in clinical practice and research is identifying and differentiating placebo from a true drug effect. A method for distinguishing between true drug and placebo response profiles in antidepressant clinical trials has been proposed (Tarpey et al. JASA 2003, 850-858). Discontinuation studies will be used to validate that method. Discontinuation trials are designed to determine the appropriate length of drug maintenance for depressed subjects who have responded to initial treatment with a drug: responders to acute treatment are randomized to continue the medication or are switched to placebo and time to relapse is assessed. The premise is that subjects who respond to the active drug will have different relapse rates depending on whether they are maintained on the drug or are switched to placebo, whereas the relapse of placebo responders will not differentiate between the two. The analysis jointly models symptoms' severity during the acute treatment and the discontinuation phases. This is accomplished by treating the estimated coefficients from the functional profiles as continuous covariates in the survival models. Refinements of this process lead to the implementation of survival analysis with a functional covariate. Data from a discontinuation study of Prozac are used to illustrate the proposed method.

email: ep120@columbia.edu

---

30. SPATIAL MODELING OF DISEASE

DETERMINANTS OF SMALL AREA RACIAL DISPARITIES IN STROKE MORTALITY IN THE SOUTHEASTERN UNITED STATES, 1999–2003

Eric C. Tassone*, Emory University
Michele Casper, Centers for Disease Control and Prevention
Lance A. Waller, Emory University
Ish Williams, Centers for Disease Control and Prevention
Katrina Moore, University of Washington and Centers for Disease Control and Prevention

In light of the increasing interest in documenting and eliminating racial and ethnic disparities in health, a pressing need exists for appropriate methods to measure these disparities at the local level. The authors measure county-level disparities in stroke mortality between African-Americans and whites in the southeastern United States during 1999-2003, for both men and women. The authors extend the typical hierarchical Bayesian disease mapping model to measure disparity using the race-gender subgroup structure of the study population. County-level maps of median disparity appear with figures displaying associated variability estimates. Race-specific maps of the burden of stroke mortality enable the reader to evaluate race-specific contributions to the disparity measure for each county. The findings highlight geographic variations across counties in magnitude of excess burden for stroke mortality borne by African Americans relative to whites. Additionally, the authors introduce covariates that measure socio-economic characteristics of the counties as possible explanatory variables.

email: etasson@sph.emory.edu

IMPACT OF PRIOR CHOICE ON LOCALIZED BAYES FACTORS FOR CLUSTER DETECTION

Ronald E. Gangnon*, University of Wisconsin–Madison

In this talk, we consider the use of a partition model to estimate regional disease rates and to detect spatial clusters. We have previously demonstrated the ability of localized Bayes factors for clustering from a model with a fixed, but overly large, number of partitions to provide useful inference about the number and locations of clusters. Here, we explore the impact of the choice of prior distribution on the potential clusters on the resulting inferences using data on breast cancer incidence in Wisconsin.

email: ronald@biostat.wisc.edu

SPATIO-TEMPORAL ANALYSIS OF EMERGENCY-ROOM VISITS FOR ISCHEMIC HEART DISEASE IN NSW, AUSTRALIA

Sandy Burden, University of Wollongong, Australia
Subharup Guha*, Harvard University
Geoff Morgan, University of Sydney, Australia
Louise Ryan, Harvard University
Ross Sparks, Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia
Linda Young, University of Florida

The recently funded Spatial Environmental Epidemiology in New South Wales project aims to use routinely collected data in NSW Australia to investigate risk factors for various chronic diseases. We present a case study focused on the relationship between social disadvantage and ischemic heart disease to highlight some of the methodological challenges that are likely to arise.

email: sguha@hsph.harvard.edu

DETECTING SPATIAL CLUSTERING IN MATCHED CASE-CONTROL STUDIES

Andrea J. Cook*, Harvard University
Yi Li , Harvard University

With the rapid development of the GIS technology, numerous methods have been proposed for detecting spatial clustering, such as the spatial scan statistic (Kulldorff, 1997), Maximized Excess Events Test (MEET) (Tango, 2000), Besag-Newell's R (Besag and Newell, 1991), k-Nearest Neighbors (Cuzick and Edwards, 1990), and the M statistic (Bonetti and Pagano, 2001). However, none have dealt with spatial data where part of the dependence is due to study design (e.g. matched pair studies). This paper proposes an extension of the spatial scan statistic that detects spatial clustering when the data are collected through a matched case-control mechanism. We further propose a robust clustering detection method, specifically, a cumulative geographic residual test that allows for discrete outcomes, matched or unmatched. Power comparisons between the spatial scan statistic and cumulative geographic residual test are made via simulations. Utilization of these methods is illustrated by a matched case-control study investigating the impact of petrochemical exposure on childhood brain and leukemia cancers.

email: acook@hsph.harvard.edu

SPATIAL ANALYSIS OF PERIODONTAL DATA USING CONDITIONALLY AUTOREGRESSIVE PRIORS HAVING TWO TYPES OF NEIGHBOR RELATIONS

Brian J. Reich*, University of Minnesota
James S. Hodges, University of Minnesota
Bradley P. Carlin, University of Minnesota

Bayesian analyses of areal data often use a conditionally autoregressive (CAR) prior distribution which allows fitted values to be smoothed toward values of neighboring regions. Sometimes it is desirable to have more than one type of neighbor relation in the spatial structure, so the different types of neighbor relations can induce different degrees of smoothing. For example, in the periodontal data sets, the degree of smoothing of neighbor pairs bridging the gap between teeth may be different from the smoothing of pairs that do not bridge such gaps. In this paper, we develop a two neighbor relation CAR (2NRCAR) model to handle this situation, and present associated theory to help explain the sometimes unusual posterior behavior of the parameters that control the different types of smoothing in this model. We also illustrate use of this model by applying it to an analysis of some periodontal data on attachment loss.

email: brianr@biostat.umn.edu

BAYESIAN CLUSTER MODELING FOR SPACE-TIME DISEASE COUNTS

Ping Yan*, University of Wisconsin–Madison
Murray K. Clayton, University of Wisconsin–Madison

Bayesian approaches to modeling spatial clustering of disease have been of great interest in recent years. However, little work has been done on modeling disease clustering in space and time. We extended the spatial cluster model developed by Gangnon & Clayton (2003, 2004) to accommodate temporal effects and space-time interactions in the general framework of generalized linear mixed effects models. MCMC methods are used in posterior inference to detect space-time clustering and to estimate relative risks. A large data set from Japan is used to illustrate these ideas.

email: yanp@stat.wisc.edu

A STATE SPACE MODEL FOR STOCHASTIC GOMPERTZ GROWTH OF CANCER TUMORS

Wai-Yuan Tan, University of Memphis
Weiming Ke*, University of Memphis

In this paper we have developed a state space model for the stochastic Gompertz growth of cancer tumors. For this state space model, the stochastic system model is derived by considering heterogeneity of cancer tumor cells involving proliferating cells (stem cells) and quiescent cells; the observation model is a statistical model based on the total number of tumor cells over time. By using this model in combination with the multi-level Gibbs sampling procedures, we have developed methods to estimate simultaneously the unknown parameters and state variables. We have applied our model and methods to some real cancer data available from the literature. Our results showed that the growth of cancer tumors can best be described by stochastic Gompertz growth through heterogeneity of cancer tumor cells. These results have some implications for predicting the growth of cancer tumors and for assessing effects of cancer therapy.

email: wke@memphis.edu

ROBUST ESTIMATION IN SEMIPARAMETRIC TRANSFORMATION MODELS FOR CENSORED POINT PROCESSES

Rajeshwari Sundaram*, University of North Carolina at Chapel Hill

Recurrent event data are frequently encountered in biomedical studies. Often the observation time in recurrent event data is censored due to loss to follow-up or administrative censoring. We consider a family of semiparametric transformation models proposed by Lin, Wei and Ying (2001) for modeling the mean function of the censored event process of interest. A class of robust minimum L_2-distance estimators for the regression parameters is proposed. This type of estimation procedure appears to be novel in the context of recurrent event data. We establish the asymptotic properties of the proposed estimators. We will also illustrate how this method can be extended to deal with marginal transformation models for multivariate recurrent events. Extensive numerical studies show that the minimum L_2-distance estimators have very good finite-sample behavior compared to existing methods. We conclude with application of our method to a cystic fibrosis trial data.

email: rsundara@uncc.edu

USE OF THE ANDERSEN-GILL MODEL TO EVALUATE TREATMENT EFFECT IN THE PRESENCE OF DISEASE PROGRESSION

Alexander C. Cambon*, University of Louisville

Recurrent Events Analysis is being increasingly used in medical research to evaluate, for example, the effect of treatment on rehospitalizations, tumor recurrences, recurrent infections, or “exacerbations”. In many cases primary interest is on assessing treatment effect. At the same time, there is also often interest is in modeling or at least taking into account disease progression. The Andersen-Gill model, using event number as a time dependent covariate, may be an alternative. However the first event in the clinical trial is often not the first event for the subject. Since the actual number of events for subjects previous to the study is often not known, the Andersen-Gill model is often not used. This presentation focuses on use of the Andersen-Gill model when other information about the subjects is known. For example, exclusion/inclusion criteria may specify that an initial event must have happened before the study in order for the subject to be included. In addition, other information gathered at the beginning of the study on each subject, such as New York Heart Association Classification, can provide additional information for each subject that may be used as a “surrogate” for this missing information.

email: accamb01@louisville.edu

SEMI-PARAMETRIC REGRESSION FOR RECURRENT EVENT DATA WITH TIME-DEPENDENT COVARIATES AND INFORMATIVE CENSORING

Xianghua Luo*, Johns Hopkins University
Mei-Cheng Wang, Johns Hopkins University

Time-dependent covariates carry useful updated information in the analysis of recurrent event data. Existing statistical methods for Cox-type regression models inherit the feature of dealing with time-dependent covariates, but they typically require independent censoring in the data collecting process. However, informative drop-out is common in follow-up studies and the independent censoring assumption is frequently violated. This paper concerns with regression analysis of time-dependent covariates for informatively censored recurrent event data. Subject- specific nonstationary Poisson processes are assumed to be the underlying model and informative censoring is characterized by a latent variable (frailty). A profile estimating function under the proportional rate regression model is proposed. Bias-correction technique through a time-transformed Poisson process is used to circumvent the estimation of the latent variable. The estimating procedures are illustrated by simulation studies and a data collected in a juvenile violent behavior study.

email: xluo@jhsph.edu

A GENERAL CLASS OF PARAMETRIC MODELS FOR RECURRENT EVENT DATA

Russell S. Stocker*, Mississippi State University
Edsel A. Pena, University of South Carolina

A general class of models for recurrent event data is presented under a fully parametric specification. An estimation scheme is given and its implementation in practice is discussed. Large sample theory is given utilizing a double time index approach. The asymptotic properties of the estimators are studied both under calendar and gap time. Finite sample properties are examined with a computer simulation study. Data on air conditioning units of a fleet of Boeing 720 jet airlines is analyzed using the class of models.

email: rstocker@math.msstate.edu

ESTIMATING THE QUALITY-OF-LIFE-ADJUSTED GAP TIME DISTRIBUTION OF SUCCESSIVE EVENTS SUBJECT TO CENSORING

Adin-Cristian Andrei*, University of Michigan
Susan Murray, University of Michigan

When studying treatment effects in the context of succesive or recurrent life events, separate analyses of the quality-of-life scores and of the inter-event (gap) times might lead to possibly contradictory conclusions. In a reconciliatory attempt, we propose a unitary and more comprehensive analysis that combines the two separate analyses by introducing the quality-of-life-adjusted gap time (QAGT) concept. Inverse probability of censoring estimators of the QAGT joint and conditional distributions are proposed and shown to be consistent and asymptotically normal. Simulations performed in a variety of scenarios indicate that the joint and conditional QAGT distribution estimators are virtually unbiased, with properly estimated standard errors and asymptotic normality features. An example from the International Breast Cancer Study Group (IBCSG) Trial V illustrates the use of the proposed estimators.

email: andreia@umich.edu

ANALYSIS OF RECURRENT EVENTS IN THE PEDIATRIC FIREARM VICTIM'S EMERGENCY DEPARTMENT VISIT DATABASE

Hyun J. Lim*, Medical College of Wisconsin
Marlene Melzer-Lange, Medical College of Wisconsin
Liu Jingxia, Medical College of Wisconsin

In many medical conditions subjects can experience recurrent or repeated events. A common feature for the recurrence time data and multi-stage failure time observations is that the events are naturally ordered and occur in a certain sequence over time. Multiple failure time models allow use of all the available information to accurately estimate the relative risk of recurrences in a given data. This study will give an insight into the choice of the best strategy according to a given recurrent event data set, in which the appropriate choice is closely related to the dependence structure. Since different methods are often required to address clinical objectives, the variety of models to be examined will provide correct analysis and a clear elucidation of recurrent events. Relevance and applicability of the proposed semi-parametric models and interpretation of the estimates of the relative risk for each recurrence in the pediatric firearm victim's emergency department visit database will be investigated.

email: hyun@mcw.edu

STATISTICAL ANALYSIS OF PANEL COUNT DATA

Do-Hwan Park*, University of Missouri
Jianguo Sun, University of Missouri
Xingqui Zhao, University of Alberta

This paper discusses statistical analysis of panel count data with focus on nonparametric comparison of point processes. Panel count data naturally occur when recurrent events are concerned. The fields in which panel count data often occur include medical follow-up studies and reliability experiments. For the problem, a class of nonparametric test statistics are proposed, which have the form of the integrated weighted differences between estimated mean functions of the point processes. The asymptotic distributions of the statistics are given and their finite sample properties are studied through simulation studies. The methodology is applied to a set of motivated panel count data from a cancer study.

email: dp41e@mizzou.edu

---

32. MISSING DATA IN LONGITUDINAL DATA ANALYSIS

LATENT CLASS REGRESSION MODELS FOR INCOMPLETE LONGITUDINAL BINARY RESPONSES

I. Dunsiger*, Brown University
Joseph W. Hogan, Brown University

Longitudinal studies in behavioral medicine are often concerned with studying patterns of behavior change. When multivariate responses are available, latent class modeling can be used to identify distinct components of a multivariate distribution, thereby providing a method for distinguishing between patterns of behavior change. Applied to longitudinal binary data, a standard assumption in latent class models is that class membership captures the marginal correlation between responses on the same individual. The current work is motivated by clinical trials in smoking cessation, where binary cessation status is measured weekly for three months. We develop a latent class regression model that allows serial correlation within class, thereby using both the mean and serial correlation to inform class membership. The model allows a separate class for responses consisting entirely of zeros. Covariate effects operate on the latent class scale. The model is used to analyze data from a three-arm longitudinal clinical trial comparing different doses of Fluoxetine (Prozac) for smoking cessation. The model identifies distinct classes of behavior change, and the treatment effects are summarized in terms of latent class membership distribution.

email: shira@stat.brown.edu

A LATENT CLASS MODEL FOR LONGITUDINAL BINARY RESPONSE DATA WITH NONIGNORABLE MISSINGNESS

Li Qin*, University of Pittsburgh
Changyu Shen, Indiana University
Lisa A. Weissfeld, University of Pittsburgh

Nonignorable missing data is a common problem in longitudinal studies. Latent class models are attractive for simplifying the modeling of missing data when the data are subject to either a monotone or intermittent missing data pattern. Roy (2003) proposed a latent class model for continuous data, in which the classes are related to the time of dropouts. In our study, we extend this approach to categorical data, dividing the observed data into two latent classes; a special class in which subjects definitely have '0' outcomes and a second one in which the outcomes can be modeled using logistic regression. In latent class models, the latent classes connect the longitudinal responses and the missingness process under the assumption of conditional independence. Thus the longitudinal responses and the missingness process are independent given the latent classes. The latent class model is also a special case of a pattern mixture model. Parameters are estimated by the method of maximum likelihood based on the above assumption and correlation between responses (le Cessie and van Houwelingen, 1994). This methodology is illustrated with a data set of weight concern in a smoking cessation study for women. For this study we compare the proposed method with a mixed effects model (Ten Have, et al., 1998) and weighted GEE (Robins et al., 1995). The results show that our method and Ten Have's model are similar and differ from the weighted GEE model. Although the results obtained using the proposed method and Ten Have's model are similar, our method is simpler to implement and can also be used for intermittent missing data.

email: liq1@pitt.edu

ROBUST METHODS FOR LONGITUDINAL DATA WITH MISSING OBSERVATIONS

Grace Yi*, University of Waterloo
Wenqing He, University of Western Ontario

Recently median regression models have received increasing attention. The models are attractive because they are robust and easy to interpret. In this paper we discuss using median regression models to deal with longitudinal data with missing observations. The inverse probability weighted generalized estimating equations approach is proposed to estimate the median parameters for incomplete longitudinal data. Consistency and the asymptotic distribution of the resultant estimators are established. The proposed method is applied to a longitudinal data set for illustration. A simulation study is conducted to assess the performance of the proposed method.

email: yyi@likelihood.math.uwaterloo.ca

A MODEL FOR INCOMPLETE LONGITUDINAL MULTIVARIATE ORDINAL DATA

Li C. Liu*, Institute of Health Research and Policy, University of Illinois at Chicago

A three-level item response theory (IRT) model is proposed for analysis of incomplete multivariate ordinal outcomes in longitudinal studies. This model accommodates missing data at any level (missing time point and/or missing item at any time point). This model allows for multiple random subject effects and the estimation of item discrimination parameters. Covariates can be at any level. Assuming either a probit or logistic response function, maximum marginal likelihood estimation (MMLE) is proposed utilizing multidimensional Gauss-Hermite quadrature for integration of the random effects. An iterative Fisher-scoring solution, which provides standard errors for all model parameters and generally converges faster than the EM algorithm, is used. A data set from a longitudinal prevention study is used to illustrate the application of the proposed model. In this study, multiple items of health behavior are repeatedly measured over time. All items are ordinal variables. Because of a planned missing design, subjects answer only two-third of all items at any time point.

email: lqi1@uic.edu

SENSITIVITY ANALYSIS AND INFORMATIVE PRIORS FOR LONGITUDINAL BINARY DATA WITH OUTCOME-RELATED DROPOUT

Joo Yeon Lee*, Brown University
Joseph W. Hogan, Brown University

Dropouts, possibly related to outcomes, are common in longitudinal data, and sensitivity analysis is indispensable to evaluate the effect of untestable assumption about dropout mechanism on study conclusion. This paper develops the pattern mixture models, composed of marginalized transition models within pattern, for repeated binary data. Within this framework we propose to introduce several approaches to systematic sensitivity analyses that will allow the analyst to explore the effects of possibly outcome-related dropout. Also, we show how to represent or convey uncertainty about common assumption such as MAR using prior distributions on the sensitivity parameters. Besides sensitivity analysis, this invites the possibility that user-supplied prior distributions that reflect beliefs about the distribution of missing responses can be incorporated into the final inferences. Methods will be illustrated using data from the OASIS study, a longitudinal clinical trial of a new intervention for smoking cessation.

email: jooyeon@stat.brown.edu

CLASSIFICATION OF MULTIVARIATE REPEATED MEASURES DATA WITH MISSING VALUES

Anuradha Roy*, University of Texas at San Antonio

The problem of classification of multivariate repeated measures data, in which the response vector is measured on each experimental unit at several time points, has become increasingly important in applications of biomedical and biological research. Missing values occur frequently in this type of data set, as observations are lost or not taken, or experimental units drop out. Traditional classification rule is not applicable in low sample size high dimensional repeated measures data, which is often the case; also it cannot be applied when the data has missing values. It cannot even incorporate the covariate information in the classification rules. We developed a new classification rule using mixed effect models, to classify multivariate repeated measures data with missing values, which can also incorporate covariate information in the classification rules. The method is illustrated with an analysis of a medical data.

email: aroy@utsa.edu

SOME PRACTICAL SOLUTIONS TO ANALYZING MESSY DATA

Monnie McGee*, Southern Methodist University

Data from real medical or field trials can be very messy, particularly if the data come from an experiment where humans are involved as subjects. This is even more problematic if the data were gathered over time. Subjects, whose numbers are small to begin with, often miss appointments, reschedule for non-evenly spaced visits, or drop out of the trial. These human problems result in incomplete and non-random missing data. Other issues include scaling of measurements among patients and difficulty in obtaining truly representative treatment groups. I show some practical and logical ways of dealing with the problems in these data, including resampling, standardization, and interpolation of missing values. I use these methods to analyze data from a small trial of the drug gabapentin in liver disease.

email: mmcgee@smu.edu

33. MULTIPLE TESTING AND FALSE DISCOVERY RATES

PROFILE SIGNIFICANCE: A POWERFUL ALTERNATIVE TO FALSE DISCOVERY RATE CONTROL

Cheng Cheng*, St. Jude Children's Research Hospital

The control of False Discovery Rate (FDR) is now a widely accepted approach to determine a statistical significance threshold for large-scale multiple tests in genome-wide studies. Although control of the level of false positive errors (false discoveries) is important, in exploratory studies such as genome-wide surveys using gene expression or SNP marker arrays to find associations with a given trait, the level of false negative errors is often considered equally important. This research addresses this by developing a significance threshold determination procedure alternative to FDR control, called "profile significance." The set of null hypotheses rejected at a given significance threshold is referred as a "rejection profile." The proposed methodology properly defines the statistical significance of a rejection profile, the "profile significance", and computes the profile significance by a permutation test. A significance threshold for the multiple tests is then determined by examining the profile significance of a large set of rejection profiles. This procedure is particularly suitable for the applications in which the test statistics are substantially dependent. Advantages and drawbacks of this methodology compared to FDR control will be presented.

email: cheng.cheng@stjude.org

FAST ESTIMATION OF SAMPLE SIZE WHILE CONTROLLING FOR FDR IN MULTIPLE TESTING

J. T. G. Hwang, Cornell University
Peng Liu*, Cornell University

Sample size estimation is important in any design of experiment and is even more so in microarray or proteomic experiments since biologists can typically afford only a few repetitions. In the multiple testing problems involving these experiments, it can be argued that it is more powerful and more reasonable to control false discovery rate (FDR) or positive FDR (pFDR) instead of type I error, e.g., family-wise error rate (FWER). See Storey and Tibshirani (2003). When controlling FDR, the traditional approach of estimating sample size by controlling type I errors is no longer applicable. Here we propose a method to calculate sample size that is applicable to controlling FDR. We generate curves for various sample sizes that can be used to find the rejection region controlling FDR in a desired level. For such a rejection region, the power is given by curves or numerical computations; hence the sample size can be easily determined for a given power. The proposed method is straightforward and instantaneous in computation, as illustrated with two sample t-tests and F-tests. The proposed method is shown based on simulation to approximate very well the power of the actual q-value procedure of Storey and Tibshirani (2003).

email: PL61@cornell.edu

PARAMETRIC AND NONPARAMETRIC FDR ESTIMATION REVISITED

Baolin Wu*, University of Minnesota
Zhong Guan, Indiana University, South Bend
Hongyu Zhao, Yale University

Nonparametric and parametric approaches have been proposed to estimate False Discovery Rate under the independent hypothesis testing assumption. The parametric approach has been shown to have better performance than the nonparametric approaches. In this article, we study the nonparametric approach and quantify the underlying relations between parametric and nonparametric approaches. Our study reveals the conservative nature of the nonparametric approach, the information loss due to the use of p-values, and establishes the connections between empirical Bayes method and p-value based nonparametric methods. Based on our results, we advocate using parametric approach, or directly modeling the test statistics instead of the p-values using the empirical Bayes method.

email: baolin@biostat.umn.edu

THE FAULTY FALSE DISCOVERY RATE: ADDRESSING BIAS IN NULL P-VALUES TO ADJUST FDR CALCULATION

Hoa Nguyen*, Carnegie Mellon University
Kathryn Roeder, Carnegie Mellon University
Larry Wasserman, Carnegie Mellon University

The false discovery procedure introduced by Benjamini and Hochberg in 1995 has become a mainstream method for large scale simultaneous inference. The procedure controls the false discovery rate (FDR) at a specified level α assuming that the distribution function F_0 of null p-values P_i is $U(0,1)$. In a recent paper, Efron (2004) brought to attention that, often, the empirical null p-values do not conform to the theoretical $U(0,1)$. Indeed, linear regression settings aimed for genome-wide association study provide good examples of a biased F_0. Under these scenarios, the number of covariates p is much greater than the sample size n, which eliminates the option of fitting the full regression model. Nevertheless, a resolution of fitting an abundant number of partial models permits an empirical estimation of the distribution of null p-values. In addressing the bias in F_0, it is more convenient to study the bias in the distribution function G_0 of z-values: $Z_i = \Phi^{-1}(P_i)$. Efron (2004) proposed a location-scale correction to the empirical distribution G_0. In this paper, we show that the bias in G_0 can not be represented by a location-scale alteration alone. We propose a skewness adaptation to G_0. We show that variants of a bias-corrected G_0 can lead to better control of FDR compared with the default $N(0,1)$. To illustrate the procedure, we examine data which are generated using a stochastic process that creates polymorphisms on chromosomal regions. The data can be analyzed using regression models.

email: htnguyen@stat.cmu.edu

P-VALUES-ONLY-BASED STEPWISE PROCEDURES FOR MULTIPLE TESTING AND THEIR OPTIMALITY PROPERTIES

Alexander Y. Gordon*, University of Rochester

A multiple testing procedure is p-values-only-based, or uninformed, if its decision on which hypotheses to reject depends only on their observed p-values, without relying on any information or assumptions about the joint distribution of p-values. We study uninformed generalized step-down procedures and associated type I error rates. We prove, under a natural condition of monotonicity, that the classic Holm procedure is the most rejective (most powerful) among all such procedures with strong control of family-wise error rate (FWER) at a given level. On the other hand, the Bonferroni procedure cannot be improved on, while preserving FWER, in the class of all monotonic generalized step-up procedures.

email: Alexander_Gordon@urmc.rochester.edu

FINITE-SAMPLE CONTROL OF THE FAMILY-WISE ERROR RATE IN MULTIPLE HYPOTHESIS TESTING

Greg DiRienzo*, Harvard University

We propose a permutation-based step-down multiple testing procedure for finite-sample control of the family-wise error rate for arbitrary data-generating distributions. Null hypotheses correspond to equality of data-generating distributions. The methods are shown to be useful in several settings, with both multiple outcome variables, which may be censored, and a large number of covariates, including (i) dimension reduction, (ii) the one-sample problem, (iii) the two-sample problem and (iv) the two- sample problem with covariates. The preference for our methods over the large-sample step-down resampling methods of van der Laan, Dudoit and Pollard (2004) in finite- samples is illustrated with a simulation study in the gene- expression context. Several real-data examples are also provided.

email: dirienzo@hsph.harvard.edu

A BAYESIAN ANALYSIS OF MULTIPLE HYPOTHESIS TESTS

Anthony L. Almudevar*, University of Rochester

A Bayesian methodology is proposed for the problem of multiple hypothesis tests for a given effect. The density of test statistics is modelled as a mixture based on hypothesis status. A full posterior measure is constructed for the mixture conditional on the observable total density. Commonly used quantities such as false discovery rates and posterior probabilities of hypothesis status can be directly calculated from the mixture, and so full posterior measures for these quantities can be directly obtained. The posterior measure is computed by sampling from a Monte Carlo Markov chain. This approach proves to be very flexible, allowing a model for the magnitude of the effects, as well as for dependence structure, to be developed and incorporated into the posterior measure. In addition, this approach is ideally suited to the situation in which the presence of large numbers of marginal, or weak, effects complicates any attempt to estimate the hypothesis mixture. In this case, a simple redefinition of the null hypothesis is proposed which makes the mixture estimation well defined and feasible.

email: anthony_almudevar@urmc.rochester.edu

---

34. COMBINING SPATIAL MEASUREMENT PARAMETERS AND DESIGN OF EXPERIMENTS TO EVALUATE THE EFFECTIVENESS OF TREATMENTS FOR PRECISION AGRICULTURE

USING SPATIAL INFORMATION IN PRECISION AGRICULTURE MANAGEMENT OF COTTON

Jeff Willers*, Mississippi State University
Chuck O'Hara, Mississippi State University
George Milliken, Kansas State University

The need for the use of spatial information in the production of Cotton is described. Natural spatial variability in soil types, elevation, and drainage patterns within fields make the currently used crop management on a whole field basis inefficient. Current technology enables the farmer to apply variable rates of fertilizer, insecticides, herbicides, and seed density. Crop harvesters are equipped with differential, global positioning system (DGPS) monitors that sample yield information at high densities (such as every second) along the harvest path across fields. Being able to characterize the spatial variability across a field and use the information to help to make decisions about how to tailor the crop management system to maximize production and profit while minimizing cost is an essential part of precision agriculture efforts. This presentation sets the stage for the next two talks which describe the process of making the spatial information usable for data analysis and the description of the design and analysis of experiments to evaluate precision agricultural management practices.

email: milliken@stat.ksu.edu

ASSEMBLING THE SPATIAL INFORMATION TO PROVIDE A DATA SET FOR STATISTICAL ANALYSES THAT CAN BE USED TO EVALUATE PRECISION

Chuck O'Hara*, Mississippi State University
Jeff Willers, Mississippi State University
George Milliken, Kansas State University

The process starts with the producer and/or subject area researcher acquisition of intricate spatial knowledge of the crop through diverse, remote sensing systems to understand the value of moving from whole field management to site specific management. The process is facilitated by tools provided by Geographic Information System (GIS) specialists who map relationships among in-field environmental conditions and crop productivity. These maps are created by combining diverse remote sensing technologies which include multispectral images, LIDAR estimates of field elevations, DGPS equipped machines that apply and record variable rate information of inputs and yield monitor data. These diverse maps and spatial data products represent compiled spatial management information throughout the field. This collection of geo-referenced data can be used with ideas of designed experiments to structure a process that determines the best combinations of treatments for the various regions of a field.

email: milliken@stat.ksu.edu

USING SPATIAL INFORMATION IN THE DESIGN AND ANALYSIS OF EXPERIMENTS USED TO EVALUATE THE EFFECTIVENESS OF PRECISION AGRICULTURE MANAGEMENT PRACTICES

George Milliken*, Kansas State University
Jeff Willers, Mississippi State University
Chuck O'Hara, Mississippi State University

The use of spatial information provided by a GIS specialist in conjunction with the subject area researcher in the design and analysis of experiments used to evaluate precision agriculture techniques is described. The studies are carried out on producer's fields and the application of the various management techniques must follow the process used by the producer. Two designs that can be used in this context are described and their analyses are discussed. The concept of experimental unit in these types of studies is discussed. The data consist of several yield monitor measures from areas within each experimental unit where the spatial characteristics of each of the areas are different. The processes are demonstrated with two examples from cotton grown on producer's fields. One important element of the success of the process is synergism of the team members.

email: milliken@stat.ksu.edu

VORONOI TESSELATION IN THE ANALYSIS OF SMALL AREA CANCER INCIDENCE AND URANIUM IN GROUND WATER

Fedele Greco*, University of Bologna, Italy
Andrew B. Lawson, University of South Carolina

Ecological regression studies are widely used in geographical epidemiology to assess the relationships between health hazard and putative risk factors. Very often health data are measured at an aggregate level because of confidentiality restrictions, while putative risk factors are measured on a different grid, i.e. independent (exposure) variable and response (counts) variable are spatially misaligned. To perform a regression of the risk on the exposure, one needs to realign the spatial support of the variables. Bayesian hierarchical models constitute a natural approach to the problem because of their ability to model the exposure field and the relationship exposure/relative risk on different levels of the hierarchy, taking proper account of the variability induced from the covariate estimation. In this paper we propose two fully Bayesian solutions to the problem. The first one is based on the kernel smoothing technique, the second one is built on the tessellation of the study region. We illustrate our methods by assessing the relationship between exposure to uranium in drinkable waters and cancer incidence in South Carolina, USA.

email: alawson@gwm.sc.edu

EFFECTS OF MODEL MISSPECIFICATION IN THE ANALYSIS OF SPATIAL DATA

Louise M. Ryan*, Harvard University

We discuss the impact of model misspecification on the analysis of areal or lattice data of the kind that arises when health outcomes are measured for geographically defined regions such as census tracts or counties. In particular, we discuss the impact on estimated covariate effects when the spatial structure of the data is either mis-modelled or ignored. We illustrate our findings with an analysis of the effects of the SEIFA index of social disadvantage on the rates of ischemic heart disease in NSW, Australia.

email: lryan@hsph.harvard.edu

THE LIMITATIONS OF SPATIAL ANALYSIS

Geoffrey M. Jacquez*, BioMedware and TerraSeer

Spatial analysis has made substantial contributions to medical geography and epidemiology, especially in the areas of exposure assessment and the identification of geographic subpopulations characterized by excesses and/or deficits of disease. However, as experimental settings geographic systems impose limitations on the knowledge that can be derived from their analysis. This presentation presents several of these limitations, including those (1) on scientific inference, (2) imposed by analytic methods and (3) emergent from static representations.

email: jacquez@biomedware.com

36. PANEL: STATISTICAL CONCERNS UNDER THE FEDERAL ADVISORY COMMITTEE ACT (FACA)

STATISTICAL CONCERNS UNDER THE FEDERAL ADVISORY COMMITTEE ACT (FACA)

Mary A. Foulkes*, OBE/CBER/FDA

Congress passed the Federal Advisory Committee Act in 1972 (Public Law 92-463), which formally recognized the merits of seeking the advice and assistance of citizens. At the same time, the Congress also sought to assure that advisory committees provide advice that is relevant, objective, and open to the public. The advisory committee process encourages public interaction with federal agencies in arriving at decisions, and members of the public are encouraged to appear before the committee during the open portions of meetings. The committees consist of individuals with recognized expertise and judgment in their field, and who have the training and experience necessary to evaluate information objectively, often under controversial circumstances. Panelists from advisory committees to several federal agencies will discuss their experiences, the process of identification and selection of committee members, communication of statistical issues within the advisory committee context across disciplines, to policy- makers and to the general public.

email: foulkes@cber.fda.gov

STATISTICAL ANALYSIS OF NOISE MODELS IN DIFFUSION TENSOR IMAGES

Hongtu Zhu*, Columbia College of Physicians and Surgeons
Dongrong Xu, Columbia College of Physicians and Surgeons
Heping Zhu, Yale University
Brady Peterson, Columbia College of Physicians and Surgeons

This paper presents a framework for use of the parametric models in constructing diffusion tensor (DT) images, and it establishes the validity of statistical inferences in diffusion tensor imaging (DTI). Estimation procedures are developed for the Rician model and two other normal models of noise. In particular, an Expectation and Maximization (EM) estimation algorithm is proposed to maximize the likelihood function of the Rician model. This paper also delineates the asymptotic distributions of four invariant measures of the diffusion tensor as well as for two measures of anisotropy. A scaled χ^2 distribution in isotropic tissue is proposed to approximate the asymptotic distribution of fractional anisotropy (FA), which provides a simple means of testing statistically whether the DT at each voxel is significantly anisotropic. In addition, we also propose a test statistic for testing singularity. Simulations characterize the bias of the estimated DTs under models for noise that are misspecified. FA has high statistical power for detecting anisotropy, while maintaining Type I error in isotropic tissues.

email: hz2114@columbia.edu

REGULARIZATION OF DIFFUSION TENSOR BRAIN IMAGES VIA THE KERNEL METHOD

Moo Chung*, University of Wisconsin-Madison

Diffusion tensor imaging (DTI) provides the directional information of water molecule diffusion in the white matter of the human brain. It is usually represented as the collection of 6 multivariate images that represent a 3 by 3 symmetric positive definite matrix called diffusion coefficients. The diffusion coefficients can be used to understand the pattern of white fibers in the brain. Most previous regularization works on diffusion tensor magnetic resonance images have been based on either anisotropic heat equations or streamline approaches. We present a novel regularization method using iterated anisotropic kernels that avoids solving diffusion equations or streamline equations while improves upon numerical stability. The bandwidth of kernel is proportionally matched to the diffusion tensor to smooth out more along the tensor fields. The proportionally constant is chosen to minimize the sum of the squared intensity difference between before and after smoothing while maximizing the smoothness of the image. The kernel method can be shown to increases the signal-to-noise ratio while preserving the anisotropy of the tensor fields. This formulation can be shown to be equivalent to the edge enhancing diffusion equation approaches.

email: mchung@stat.wisc.edu

EIGENANALYSIS OF DTI TENSORS ACROSS POPULATIONS

Jonathan Taylor*, Stanford University

Many voxel-based morphometry(VBM) studies of DTI data use the fractional anisotropy (FA) to convert the multivariate data to univariate data. This renders the data suitable for standard VBM analysis. FA is a normalized measure of the dispersion in the eigenvalues of the DTI tensor which contains no directional information. In this talk, we present some multivariate approaches based on the full eigenstructure of the DTI tensors, that is, both the eigenvalues and eigenvectors. Our approach, in theory, allows us to distinguish regions where the eigenvectors differ across populations from regions where the eigenvalues differ across populations. We illustrate our technique on a data set comparing dyslexic subjects to controls.

e-mail: jonathan.taylor@stanford.edu

RANDOM FIELDS OF MULTIVARIATE TEST STATISTICS, WITH AN APPLICATION TO SHAPE ANALYSIS

Keith Worsley*, McGill University

Our data are random fields of multivariate normal observations, and we fit a multivariate linear model with common design matrix at each point. We are interested in detecting those points where some of the coefficients are non-zero using classical multivariate statistics evaluated at each point. The problem is to find the P-value of the maximum of such a random field of test statistics. We approximate this by the expected Euler characteristic of the excursion set. Our main result is a very simple method for calculating this, which not only gives us the previous result of Cao & Worsley (1999a) for Hotelling's T^2, but also random fields of Roy's maximum root, maximum canonical correlations (Cao & Worsley, 1999b), bar-chi^2 (Lin & Lindsay, 1997; Takemura & Kuriki, 1997), and multilinear forms (Kuriki & Takemura, 2001) in general. The trick involves approaching the problem from the point of view of Roy's union-intersection principle. The results are applied to a problem in shape analysis, where we look for brain damage due to non-missile trauma.

e-mail: worsley@math.mcgill.ca

38. RECENT DEVELOPMENTS ON METHODS FOR MULTIVARIATE FAILURE TIME DATA

RECENT AND NEEDED DEVELOPMENTS IN THE ANALYSIS OF CORRELATED FAILURE TIME DATA

Ross Prentice*, Fred Hutchinson Cancer Research Center and University of Washington

A brief overview will be provided of methods for the regression analysis of marginal hazard rates and pairwise dependencies, and for joint survivor function estimation more generally, based on censored correlated failure time data. A description of further needed developments with particular reference to the auxiliary and surrogate variable problems will be described.

email: rprentic@fhcrc.org

SOME APPROACHES TO MULTIVARIATE FAILURE TIME ANALYSIS

Jerry Lawless*, University of Waterloo

This talk will consider approaches to the regression analysis of multivariate failure times, based on extensions of Cox (proportional hazards) and accelerated failure time (log-location-scale) models to multivariate settings. Pros and cons of different approaches with respect to features such as efficiency, interpretability, robustness, the ability to handle multilevel association, and the ability to deal with different types of censoring will be discussed and illustrated.

email: jlawless@uwaterloo.ca

MODEL SELECTION FOR MULTIVARIATE SURVIVAL DATA ANALYSIS

Runze Li*, Penn State University
Jianwen Cai, University of North Carolina at Chapel Hill
Haibo Zhou, University of North Carolina at Chapel Hill
Jianqing Fan, Princeton University

In this talk, I will present a penalized pseudo-partial likelihood method for variable selection to multivariate failure time data with a growing number of regression coefficients. Under certain regularity conditions, we show the consistency and asymptotic normality of the penalized likelihood estimators. We further demonstrate that, for certain penalty function with proper choices of regularization parameters, the resulting estimator can correctly identify the true model, as if it were known in advance. Based on a simple approximation of the penalty function, the proposed method can be easily carried out with Newton-Raphson algorithm. We conduct extensive Monte Carlo simulation studies to assess the finite sample performance of the proposed procedures. We illustrate the proposed method by analysing a dataset from the Framingham Heart Study.

email: rli@stat.psu.edu

BAYESIAN FRAILTY MODELS BASED ON BOX-COX TRANSFORMED HAZARDS

Guosheng Yin*, University of Texas M. D. Anderson Cancer Center
Joseph G. Ibrahim, University of North Carolina at Chapel Hill

Due to natural or artificial clustering, multivariate failure time data often arise in biomedical research. To account for the intracluster correlation, we propose a novel class of frailty models by imposing the Box-Cox transformation on the hazard functions. This class of models generalizes the relationships between the baseline hazard and the hazard functions, which includes the proportional and the additive hazards frailty models as two special cases. Since hazards cannot be negative, complex multidimensional nonlinear parameter constraints must be imposed in the model formulation. To facilitate a tractable computational algorithm, the joint priors are constructed through a conditional-marginal specification. The conditional distribution of the prior specification is univariate and absorbs the parameter constraints, while the marginal part is free of constraints. We propose a Markov chain Monte Carlo (MCMC) computational scheme for sampling from the posterior distribution of the parameters. We derive an MCMC approximation for the conditional predictive ordinate to assess model adequacy, and illustrate the proposed method with a dataset.

email: gsyin@mdanderson.org

39. ADVANCES IN APPLICATIONS OF LATENT VARIABLE MODEL ANALYSIS FOR HEALTH SERVICE RESEARCH

OPTIMAL DESIGN FOR STUDIES WITH MULTIVARIATE OUTCOMES

Chen-Pin Wang*, University of Texas Health Science Center at San Antonio

Maximizing efficiency, information, or utility often is the ultimate goal of study design. In health services research, a more realistic goal, however, is to achieve optimality under feasible scenarios. Drawing from Keifer (1959) and the literature that follows, optimal designs subject to constraints, with respect to efficiency and information criterion, are derived for studies involving multivariate outcomes arising from latent variable models. Following the same principle, two optimality criteria derived from the Bayesian perspective are examined, which are the inverse of the (co)variance matrix and the Kullback-Leibler distance associated with the posteriors. Asymptotic properties of the four optimality criteria are derived. It is shown that the criterion based on the Kullback-Leibler distance is superior when the assumed model is mis-specified.

email: cwang@verdict.uthscsa.edu

MODELING HETEROGENEITY IN A RANDOMIZED CLINICAL TRIAL FOR TREATMENT OF ALCOHOL DEPENDENCE

Jennie Z. Ma*, University of Texas Health Science Center at San Antonio
Chen-Pin Wang, University of Texas Health Science Center at San Antonio

Extending from the two-part latent growth mixture modeling technique, a new model is proposed to examine the variation of treatment effects with better precision - variation across different trajectory classes as well as over time (characterized by change-points). A randomized longitudinal clinical trial on Topiramate, a treatment for alcohol dependence and craving among heavy drinkers is used to demonstrate how to determine the sources of latent heterogeneity using the pseudoclass diagnostic technique in conjunction with cum-sum statistics.

email: maj2@uthscsa.edu

CAUSAL INFERENCE FOR LATENT SUBPOPULATIONS

Booil Jo*, Stanford University

Understanding differential treatment efficacy across subpopulations may help researchers to improve treatment plans and to better design future studies. Given complex structures of heterogeneity information involving missing (latent) data, latent class and normal mixture models are often adopted to maximize the use of available information in the classification of individuals. Although flexible, efficacy estimation based on latent classes can be arbitrary and subjective due to the exploratory nature of latent variable approaches. Some modeling possibilities to obtain causal inference for latent subpopulations will be discussed.

email: booil@stanford.edu

LATENT VARIABLE MODELS FOR ASSESSING TREATMENT COST-EFFECTIVENESS

Alka Indurkhya*, Harvard University

While latent variable models have become quite popular to assess effectiveness of treatments in clinical and epidemiologic studies, they are rarely used in modeling costs. The main focus of this paper will be to present the advantages of jointly modeling costs and effectiveness using latent variable models. Assumptions necessary to permit such a joint modeling will be presented along with data illustrations of issues in estimation and modeling that arise when these assumptions are violated.

email: aindurkh@hsph.harvard.edu

MEDIATION ANALYSES WITH STRUCTURAL MEAN MODELS

Tom Tenhave*, University of Pennsylvania School of Medicine
Marshall Joffe, University of Pennsylvania School of Medicine

We present a linear structural mean model approach for analyzing mediation of a randomized baseline intervention's effect on a univariate follow-up outcome. Unlike standard mediation analyses, our approach does not assume that the mediating factor is randomly assigned to individuals by taking advantage of baseline randomization. The proposed G-estimation procedure represents an extension of the work on treatment non-adherence by Robins and Greenland (1992) and Ten Have et al. (2004) to estimation of direct and indirect effects of a randomized baseline factor. Simulations show good test and confidence interval performance and robustness under unmeasured confounding, in contrast to standard mediation approaches. We also present results from analyzing how anti-depressant medication mediates an intent-to-treat effect of a randomized encouragement intervention on follow-up Hamilton depression scores.

email: ttenhave@cceb.upenn.edu

---

40. CATEGORICAL DATA ANALYSIS AND EXPERIMENTAL DESIGN

NONLINEAR TEST FOR CATEGORICAL DATA

Momiao Xiong*, University of Texas Health Science Center at San Antonio

A χ^2 test is a standard statistical method for analysis of categorical data. However, a χ^2 test is a linear test, which is based on a linear function of observed variables. Theoretical analysis shows that any linear transformation of observed variables will not change test statistics. To amplify the ratio of "signal" over noise implied in the data and to overcome limitations of linear test, we propose to develop nonlinear tests for analysis of categorical data. In this report, we present a general framework for nonlinear tests and develop statistics for nonlinear tests. We will study distributions of nonlinear test statistics under the null hypothesis and develop analytic tools for power calculations. Several nonlinear tests such as entropy-based, exponential and polynomial tests will be investigated and their power will be compared. Applications of nonlinear tests to genetic association studies of complex diseases will be discussed. Two real data sets from testing association of a COMT haplotype with schizophrenia and association of the MMP-2 gene with esophageal carcinoma are used to evaluate the performance of the proposed nonlinear tests for association studies. The results show that the nonlinear tests obtain much smaller p values than the linear tests.

email: mxiong@sph.uth.tmc.edu

EVALUATION CRITERIA FOR DISCRETE CONFIDENCE INTERVALS: BEYOND COVERAGE AND LENGTH

Paul W. Vos, East Carolina University
Suzanne S. Hudson*, East Carolina University

Confidence intervals for discrete distributions are often evaluated only by coverage and average length. We discuss two additional critera, p-confidence and p-bias. The choice of these criteria is motivated by the interpretation of a confidence interval as being the set of parameter values not rejected by a hypothesis test. Using these additional criteria we compare a number of equal-tailed confidence intervals for the binomial distribution. It is shown that methods that produce superior intervals, as measured by coverage and length, need not perform well in terms of p-confidence and p-bias. Cox's measuring device example is discussed to motivate the need for criteria beyond coverage and length.

email: hudsons@mail.ecu.edu

ON HYPOTHESIS TESTING AND CONFIDENCE INTERVALS FOR THE DIFFERENCE OF TWO INDEPENDENT POISSON RATES

William W. B. Wang*, Merck Research Laboratories
Jin Xu, Merck Research Laboratories
Santosh Sutradhar, Merck Research Laboratories
Ivan S. F. Chan, Merck Research Laboratories

In controlled clinical trials with long term follow-up and/or staggered enrollment, the disease risk is often characterized by the incidence rate based on person-time data (i.e., number of cases per 1000 person-years). The difference of incidence rates between the test and control groups are sometimes used to measure the treatment effect. The reciprocal of this difference can be readily interpreted as "number of person-years needed to treat in order to cure or prevent 1 case of disease," an interpretation that is appealing to health care policy makers or physicians. In this presentation, we investigate multiple methods for hypothesis testing and for computing test-based confidence intervals for the difference of two independent Poisson rates. These methods are: 1) an asymptotic Z-type method based on variances estimated from the observed incidence rates; 2) an asymptotic Z-type method based on variances estimated from the constraint MLEs (Miettinen and Nurminen, Stat. Med. 1985, 4, 213-226); and 3) an exact method conditional on the total number of cases in the trial. These methods are extended and discussed in the setting of stratified clinical trials. Simulation studies are performed to evaluate their properties in terms of size, power and coverage probability. Numerical examples are provided for illustrations.

email: william_wang@merck.com

MULTIPLE COMPARISONS FOR ODDS RATIOS

Melinda H. McCann*, Oklahoma State University
Joshua M. Tebbs, Kansas State University

When reporting odds ratios for a categorical variable with more than two levels, it is common to provide confidence intervals involving a reference level with each of the remaining levels. Such intervals are typically computed using techniques which specify a fixed error rate for each interval. Unfortunately, this does not allow the researcher to draw simultaneous conclusions among the different levels because the familywise error rate in the set of comparisons is not controlled. To address this issue, we derive simultaneous procedures for odds ratios by (i) exploiting their relationship to parameters in a logistic regression model, and (ii) utilizing large-sample methods based on multinomial counts in a contingency table. We evaluate the performance of both approaches and illustrate our methods using data from a medical study investigating the effects of preeclampsia.

email: tebbs@ksu.edu

ORTHOGONAL ARRAYS OF 2- AND 3-LEVELS FOR LEAN DESIGNS

Changxing Ma*, University of Florida
Ling-Yau Chan, University of Hong Kong

When an orthogonal array (OA) of n rows is used as the design matrix in an experiment, n is the number of runs. In an OA of q levels, n is an integer multiple of $q^2 $.In an experiment, if the number of runs cannot be set exactly equal to the number of rows of an OA because of constraints in resources or other reasons, the experimenter may use a design matrix formed by omitting some rows of an OA. If such a design matrix is used, the number of observed response obtained may not be enough for estimation of all the effects corresponding to columns of the orthogonal array. A lean design is a design matrix formed by deleting some rows and columns of an OA, which still allows efficient estimation the effects of the factors corresponding to the remaining columns of the OA. In this article, the authors discuss lean designs of 2 and 3 levels, and provide D-optimal OA's from which lean designs can be formed.

email: cma@biostat.ufl.edu

POWER CALCULATIONS FOR A ZERO-INFLATED POISSON MODEL

John M. Williamson, Centers for Disease Control and Prevention
Hung-Mo Lin*, Pennsylvania State University College of Medicine
Allen W. Hightower, Centers for Disease Control and Prevention

We propose power calculations for a two-sample test using a zero-inflated Poisson model. We calculate the power based on a Wald test for a two-group comparison where the group covariate can be used in the modeling of both the count data and the 'excess zero' data, or in either part separately. The resulting Wald test statistic has a non-central distribution. We present simulations to detail the performance of the proposed power calculations. Analyses of two biomedical studies are used for illustration.

email: hlin@psu.edu

DOSE-ADJUSTED MANTEL-HAENSZEL TESTS FOR NUMERIC SCALED STRATA

Stuart A. Gansky*, University of California, San Francisco

A recent three arm parallel groups randomized clinical trial had a protocol deviation causing participants to have fewer active doses of an in-office treatment than planned. Thus, the statistical analysis plan was modified from a minimal assumption randomization-based extended Mantel-Haenszel (EMH) test and a dose-adjusted EMH (DAEMH) test was developed wtih an extra set of weights corresponding to the number of active doses. A set of Monte Carlo simulations was undertaken with and without actual dose-response effects and 1000 replicates to estimate empirical size and power. Results showed improved power for the DAEMH versus the EMH for an actual dose-response and not much difference in power when an actual dose-response did not exist. Support: US DHHS/NIH/NIDCR, NCMHD U54 DE 14251

email: sgansky@itsa.ucsf.edu

ADAPTIVE DESIGN AND MULTIPLE COMPARISON ADJUSTMENT IN MULTIPLE DOSE CLINICAL TRIALS

Liji Shen*, Sanofi-Aventis

Stepwise Over-Correction (SOC) is a new method to evaluate drug effect in a multiple dose clinical trial. The author found that the difficulty of statistical inference in multiple-dose trials was caused by improper statistical estimates for the parameter in the tests. For example, the maximum observed response rate overestimates the response rate of the selected dose. The new method (SOC) was proposed to correct the bias due to the overestimation. Simulations show that SOC is better than methods such as Bonferroni's procedure. This new method is particularly useful in a two-stage phase II/III design that contains a dose selection stage and a confirmatory stage.

email: liji.shen@sanofi-synthelabo.com

SAMPLE SIZE ESTIMATION AND DESIGN SELECTION FOR A RANDOMIZED TRIAL SUBJECT TO INFORMATIVE DROPOUTS

Wenjun Li*, University of Massachusetts Medical School

In a randomized trial on osteoarthritis patients, the outcome is change in knee cartilage volume measured by magnetic resonance image scans (MRIs). The trial cost is largely driven by the number of required MRIs. Such trials also suffer dropouts that depend on the outcome. This paper illustrates a method for estimating sample size and selecting a design that minimizes the cost while accounting for informative dropouts. First, the data are simulated by assuming a first-order autoregressive process and using the expected study effects and parameters from published trials. The dropout process is simulated using a logistic function depending on patients' cartilage volume and characteristics, and observations marked as dropout are excluded from analysis. The treatment effect is then tested using a linear mixed model weighted by the inverse of the estimated probabilities of subjects remaining in the trial at each time point. Various competing designs are simulated by varying trial size and duration and number of MRIs per patient. For each design, the process is repeated 2,500 times, and the estimated power is the proportion of the correct rejection of the null hypothesis at the specified significance level. Optimal design can be determined based on the expected number of MRIs while satisfying power requirement.

email: wenjun.li@umassmed.edu

AN ADAPTIVE DOSE-FINDING DESIGN INCORPORATING BOTH TOXICITY AND EFFICACY

Wei Zhang*, University of Iowa
Daniel J. Sargent, Mayo Clinic
Sumithra Mandrekar, Mayo Clinic

Novel therapies are challenging the standards of drug development. Agents with specific biologic targets and limited toxicity require novel designs to determine doses to be taken forward into larger studies. In this paper, we describe an approach that incorporates both toxicity and efficacy data into the estimation of the biologically optimal dose of an agent in a phase I trial. The approach is based on the flexible continuation-ratio model, and uses straightforward optimal dose selection criteria. Dose selection is based on all patients treated up until that time point, using a continual reassessment method approach. Our simulation studies demonstrate that the proposed design, which we call TriCRM, has favorable operating characteristics compared to other designs previously proposed and in use.

email: wei-zhang-1@uiowa.edu

SAMPLE SIZE COMPUTATION FOR MULTIVARIATE OUTCOMES

Peng Huang*, Medical University of South Carolina
Barbara C. Tilley, Medical University of South Carolina
Yuko Palesch, Medical University of South Carolina
Jordan Elm, Medical University of South Carolina

Sample size computation is required for designing a clinical trial. For a trial with multiple primary outcomes, often the sample size is computed for each outcome using some adjustment of the overall Type I error, such as the Bonferroni correction, and the largest sample size required for a single outcome is selected for the trial. Many of these approaches, in particular the Bonferroni method and its extensions, can be too conservative when the interest is to detect a consistent but small improvement over a large number of outcomes. Thus, rather than conducting multiple tests on the primary outcomes of interest separately, a global statistical test, such as O'Brien's rank-sum test, can be adopted. This statistical approach requires a smaller sample size than the sample size computed using the Bonferroni approach. To compute sample size for O'Brien's rank-sum test for non-survival outcomes and logrank test for survival outcomes we propose a new sample size computation formula. Applications in Parkinson's disease clinical trials are presented.

email: huangp@musc.edu

THE VALUE OF INFORMATION AND OPTIMAL CLINICAL TRIAL DESIGN

Andrew R. Willan*, Hospital for Sick Children
Eleanor M. Pinto, University of Toronto

Traditional sample size calculations for randomized clinical trials depend on arbitrarily chosen factors, such as type I and II errors. Type I error, the probability of rejecting the null hypothesis of no difference when it is true, is most often set to 0.05, regardless of the cost of such an error. In addition, the traditional use of 0.2 for the type II error means that the money and effort spent on the trial will be wasted 20% of the time even when the true treatment difference is equal to the smallest clinically important one and, again, will not reflect the cost of making such an error. A pragmatic trial (otherwise know as an effectiveness trial or management trial) is essentially an effort to inform decision-making, i.e. should Treatment be adopted over Standard? As such, a decision theoretic approach will lead to a more optimal sample size determination. Using incremental net benefit and the theory of the expected value of information, and taking a societal perspective, one can determine the sample size that maximizes the difference between the cost of doing the trial and the value of the information gained from the results. The methods are illustrated using examples from oncology and obstetrics.

email: andy@andywillan.com

BIVARIATE DESIGNS IN PHASE II TRIALS

Menggang Yu*, Indiana University School of Medicine
Constantin Yiannoutsos, Indiana University School of Medicine

In this talk, we consider study designs in phase II clinical trials where both efficacy and toxicity are primary outcomes. One stage and two stage designs will be discussed and a real example will be given. A new design which allows for trade-off between efficacy and toxicity is proposed. Related computational aspects of the design will also be presented.

email: meyu@iupui.edu

A TWO-STAGE SAMPLE SIZE RECALCULATION PROCEDURE FOR PLACEBO- AND ACTIVE-CONTROLLED NON-INFERIORITY TRIALS

Todd A. Schwartz*, University of North Carolina at Chapel Hill
Jonathan S. Denne, Eli Lilly and Company

Many non-inferiority trials of a test treatment versus an active control may also, if ethical, incorporate a placebo arm. Inclusion of a placebo arm enables a direct assessment of assay sensitivity. It also allows construction of a non-inferiority test that avoids the problematic specification of an absolute non-inferiority margin, and instead evaluates whether the test treatment preserves a pre-specified proportion of the effect of the active control over placebo. We describe a two-stage procedure for sample size recalculation in such a setting that maintains the desired power more closely than a fixed sample approach when the magnitude of the effect of the active control differs from that anticipated. We derive an allocation rule for randomization under which the procedure preserves the Type I error rate, and show that this coincides with that previously presented for optimal allocation of the sample size among the three treatment arms.

email: tschwart@bios.unc.edu

---

42. COMPETING RISKS AND CURE RATES

A STUDY OF INVERSE PROBABILITY OF CENSORING WEIGHTED ESTIMATORS OF CUMULATIVE INCIDENCE FUNCTION FOR COMPETING RISKS DATA

Xu Zhang*, Medical College of Wisconsin
Meijie Zhang, Medical College of Wisconsin

Cumulative incidence curves are important summary curves in analyzing competing risks data. In this paper we consider three nonparametric estimators of cumulative incidence function: the usual estimator based on cause-specific hazards, the product-limit estimator based on a subdistribution hazard (Fine and Gray, 1999), and the weighted empirical estimator. We show that these three estimators are identical. Based on these estimators, different variance estimators can be derived. We examine the performances of the variance estimators through simulation. By using covariate adjusted weights, we propose an improved estimator which is consistent in case of dependent censoring, if certain assumptions hold. The performance of the improved estimator is also examined through simulation.

email: xzhang@mcw.edu

NONPARAMETRIC ESTIMATION WITH LEFT TRUNCATED SEMI-COMPETING RISKS DATA

Limin Peng*, University of Wisconsin–Madison
Jason P. Fine, University of Wisconsin–Madison

Cause-specific hazard and cumulative incidence function are of practical importance in competing risks studies. Inferential procedures for these quantities are well developed and can be applied to semi-competing risks data, where a terminating event censors a non-terminating event, after coercing the data into the competing risks format. Complications arise when there is left truncation of the terminating event, as often occurs in observational studies. The competing risks analysis naively truncates the non-terminating event using the left truncation time for the terminating event, which may lead to large efficiency losses. We propose simple nonparametric estimators which use all semi-competing risks information and do not require aritificial truncation. The uniform consistency and weak convergence of the estimators are established and variance estimators are provided. Simulation studies and an analysis of a diabetes registry demonstrate large efficiency gains over the naive estimators.

email: pengl@stat.wisc.edu

COMPETING RISK TRANSFORMATION MODELS WITH MISSINGNESS

Guozhi Gao*, North Carolina State University
Anastasios A. Tsiatis, North Carolina State University

We consider the problem of estimating the regression coefficients in a competing risks model, where the relationship between the cause-specific hazard for the cause of interest and covariates is described using linear transformation models, and when cause of failure is missing at random for a subset of individuals. Using the theory of Robins et al. (1994) for missing data problems and the approach of Chen et al. (2002) for estimating regression coefficients for linear transformation models, we derive augmented inverse probability weighted complete-case estimators for the regression coefficients that are doubly-robust, in the sense that the estimators are consistent and asymptotically normal when either the model for the probability of missingness or the model for the probability of cause of failure being the cause of interest are correctly specified. A convenient computational algorithm is also provided for obtaining the estimates. Simulation results demonstrated the adequacy of the asymptotic theory of our estimator, and the superiority (less bias and higher efficiency) of this estimator to other existing estimators for two important cases of linear transformation models: the proportional hazards model and the proportional odds model.

email: ggao@ncsu.edu

COMPETING MODELS FOR COMPETING RISK

Jack Kalbfleisch, University of Michigan
Yining Ye*, University of Michigan

In many studies, survival data involve several types of failure. When the effects of covariates of interest differ markedly for the different types, a type-specific analysis is important. Different approaches have been proposed to model competing risk data. One is to model directly the cause-specific hazard function. Fine and Gray (1999) suggested modeling the type-specific subdistribution (or cumulative incidence) function, which has the advantage of directly assessing covariate effects on this function. We compare and contrast the two modeling strategies using Cox type models, and demonstrate that the subdistribution hazards proposed by Fine and Gray cannot be modeled independently and consistently over all types of failure. We also consider the censoring issues in the new modeling strategy, which are much more complicated than the cause- specific hazard approach. In a series of examples, we find that the cause-specific hazard models provide a simpler and clearer picture of the situation, and at least in these examples seem to provide a better approach to competing risk data analysis. Bootstrapping methods can be used to make inferences about the subdistribution function.

email: yey@umich.edu

FLEXIBLE CURE RATE MODELING UNDER LATENT ACTIVATION SCHEMES

Freda W. Cooner*, University of Minnesota
Sudipto Banerjee, University of Minnesota
Bradley P. Carlin, University of Minnesota
Debajyoti Sinha, Medical University of South Carolina

Survival models have been and continue to be extremely popular in analyzing time-to-event data. A particular class of survival models, called cure rate models, tries to estimate the probability of subjects getting cured. Research articles on cure rate models can be traced back to over four decades, but recent developments in formulating cure models in Bayesian hierarchical settings have evoked much discussion, particularly with regard to relationships and preferences between the different types of cure models. Broadly speaking, this talk proposes a unifying class of cure models that include existing classes as special cases, while allowing new and flexible model-building. In the process, the relationship between classical and hierarchical cure models is also clarified. Issues such as regressing on the cure fraction and associated propriety issues are also addressed. The models are illustrated with a melanoma data set and a breast cancer data set, revealing possibly different underlying mechanisms that lead to cure.

email: xiyunwu@biostat.umn.edu

A NEW APPROACH TO TESTING FOR SUFFICIENT FOLLOW-UP IN CURE-RATE ANALYSIS

Lev B. Klebanov, Charls University, Praha, Czech Republic
Andrei Y. Yakovlev*, University of Rochester

The problem of sufficient follow-up arises naturally in the context of cure rate estimation. This problem was brought to the fore by Maller and Zhou (1992, 1994) in an effort to develop nonparametric statistical inference based on a binary mixture model. The authors proposed a statistical test to help practitioners decide whether or not the period of observation has been long enough for this inference to be theoretically sound. The test is inextricably entwined with estimation of the cure probability by the Kaplan-Meier estimator at the point of last observation. While intuitively compelling, the test by Maller and Zhou does not provide a satisfactory solution to the problem because of its unstable and non- monotonic behavior when the duration of follow-up increases. The present paper introduces an alternative concept of sufficient follow-up allowing derivation of a lower bound for the expected proportion of immune subjects in a wide class of cure models. By building on the proposed bound, a new statistical test is designed to address the issue of the presence of immunes in the study population. The usefulness of the proposed approach is illustrated with an application to survival data on breast cancer patients identified through the NCI Surveillance, Epidemiology and End Results Database.

email: andrei_yakovlev@urmc.rochester.edu

43. ANALYZING MICROARRAY DATA

COMPARISON OF NORMALIZATION TECHNIQUES FOR CDNA MICROARRAY DATA

Kimberly F. Sellers*, University of Pennsylvania
Jeffrey C. Miecznikowski, Carnegie Mellon University
William F. Eddy, Carnegie Mellon University

Microarray technology has been known to contain systematic variation as a result of its image and data processing procedures. Various normalization techniques have been proposed to address this issue. We present a means for comparing some of these proposed methods (via the statistical package R), in consideration of how these models normalize the data and affect the detection of differential expression.

email: ksellers@cceb.upenn.edu

A NON-PARAMETRIC APPROACH OF GENE SELECTION IN OLIGONUCLEOTIDE ARRAYS

Dung-Tsa Chen*, University of Alabama at Birmingham
James Chen, FDA/NCTR
Chen-An Tsai, University of Alabama at Birmingham
Seng-jaw Soong, University of Alabama at Birmingham

Affymetrix oligonucleotide arrays have a two-layer data structure with probe information in lower level and gene information in the upper level. Conventional approaches often summarize probe-level information into gene level expression so that standard statistic methods can be directly applied for gene selection. Though these approaches reduce data dimension in a manageable scale, valuable probe level information, such as probe effect and interaction effect between probe affinity and treatment effect, may be lost by such strategy. Thus, the use of gene level data may not be optimal for data analysis. In this study, we propose a non-parametric approach to analyze probe level gene data. We use a rank approach to normalize probe intensities and a two-stage testing procedure to select altered genes.

email: dtchen@uab.edu

INCORPORATING MULTIPLE CDNA MICROARRAY SLIDE SCANS—APPLICATION TO SOMATIC EMBRYOGENESIS IN MAIZE

Tanzy M. Love*, Iowa State University
Alicia L. Carriquiry, Iowa State University

Microarray data are subject to multiple sources of measurement error. One source of potentially significant error is the settings of the instruments (laser and sensor) that are used to obtain the measurements of gene expression. Because `optimal' settings may vary from slide to slide, operators typically scan each slide multiple times and then choose the reading with the fewest over- exposed and under-exposed spots. We propose a hierarchical modeling approach to estimating gene expression that combines all available readings on each spot. The basic premise is that all readings contribute some information about gene expression and that after appropriate re- scaling, it would be possible to combine all readings into a single estimate. We illustrate the use of the model using expression data from a maize embryogenesis experiment and assess the statistical properties of the proposed expression estimates using a simulation experiment. As expected, combining all available scans using a reasonable approach to do so results in expression estimates with noticeably lower bias and root mean squared error relative to other approaches that have been proposed.

email: tanzy@iastate.edu

STATISTICAL DEVELOPMENT AND EVALUATION OF MICROARRAY DATA FILTERS

Stanley B. Pounds*, St. Jude Children's Research Hospital
Cheng Cheng, St. Jude Children's Research Hospital

Filtering is a common practice used to simplify the analysis of microarray data by removing from subsequent consideration probe sets believed to be unexpressed. The m/n filter, which is widely used in the analysis of Affymetrix data, removes all probe sets having fewer than m present calls among a set of n chips. Two alternative filters, the pooled p-value filter and the error minimizing pooled p-value filter, are proposed. The pooled p-value filter combines information from the present/absent p-values into a single summary p-value. The error minimizing pooled p-value filter compares the summary p-value with a threshold determined to minimize a total error criterion. We show that the pooled p-value filter is the uniformly most powerful statistical test under a reasonable beta model and that it exhibits greater power than the m/n filter in all scenarios considered in a simulation study. The pooled p-value and error-minimizing pooled p-value filters clearly perform better than the m/n filter in a case-study analysis. A filter impact analysis shows that the use of even the best filter may actually hinder the ability to discover interesting probe sets or genes.

email: stanley.pounds@stjude.org

LINEAR MIXED EFFECTS MODELS FOR DUAL COLOR MICROARRAY INTENSITY RATIOS

Guilherme J. M. Rosa*, Michigan State University
Juan Pedro Steibel, Michigan State University
Robert J. Tempelman, Michigan State University

Linear models have been widely used for comparing gene expression profiles across groups or populations (within experimental or observational settings) in dual color microarray experiments, such as with cDNA or long oligonucleotide platforms. Fixed effects models have been suggested for modeling either log fluorescence intensities (Kerr and Churchill, Genet. Res. 77: 123-128, 2001) or their differences (log ratios) for each spot on an array (Yang and Speed, Nat. Rev. Genet. 3: 579-588, 2002). Fixed effects models, however, do not easily facilitate modeling of various sources of correlation between observations, particularly the partitioning of technical from biological levels of replication. Extreme care should then be taken with the definition of the experimental unit in order to appropriately specify experimental error (Churchill, Nature Genet. 32: 490-495, 2002). To overcome these difficulties, linear mixed effects models have been suggested for modeling log fluorescence intensities (Wolfinger et al., J. Comp. Biol. 8: 625-637, 2001). However, fixed effects models are still commonly used for modeling log ratios in two-channel microarray systems. In addition, the models currently suggested in the literature for log ratio intensities ignore the possibility of gene x dye interactions. We present alternative mixed effects models for the analysis of log ratio intensities in different experimental layouts (including dye-swap and incomplete block structures). Similarities and differences between both mixed effects models (i.e. for log fluorescence intensities or their log ratios) are highlighted. With simple examples, it is shown how linear models for log intensity ratios should be specified to deal with general covariance structures and gene specific dye effects.

email: rosag@msu.edu

PROBE-LEVEL CORRECTION AND ANALYSIS OF AFFYMETRIX GENECHIPS

Fenghai Duan*, Yale University School of Medicine
Heping Zhang, Yale University School of Medicine

For certain Affymetrix GeneChips, it is necessary to correct the texture effect on the images before doing the common background subtraction and normalization. In this study, we explore a way to assess and correct the texture effect. Without this correction, we found that the texture effect affected the performance of the conventional approach in identification of differentially expressed genes. Furthermore, we proposed a new approach to summarize and analyze the probe-level data and found that it can help remove the local artifacts from the residual images, which are commonly associated with conventional approaches. Finally, we also tried to evaluate and adjust for the effects of several other biological factors on the probe intensity.

email: f.duan@yale.edu

A TEST STATISTIC FOR TESTING TWO-SAMPLE HYPOTHESES IN MICROARRAY DATA ANALYSIS

Lev Klebanov, Charls University, Praha, Czech Republic
Alexander Gordon, University of Rochester
Yuanhui Xiao*, University of Rochester
Hartmut Land, University of Rochester
Andrei Yakovlev, University of Rochester

We introduce a test statistic intended for use in nonparametric testing of the two-sample hypothesis with the aid of resampling techniques. This statistic is constructed as an empirical counterpart of a certain distance measure, N, between the distributions F and G from which the samples under study are drawn. The distance measure N can be shown to be a probability metric. For the N-statistic, it can be shown that a specific resampling procedure (resampling analog of permutations) provides a rational way of modeling the null distribution whenever the underlying null hypothesis is composite. More specifically, this procedure mimics the sampling from a null distribution which is, in some sense, the ``least favorable'' for rejection of the null hypothesis. No statement of such generality can be made for the t- statistic. The usefulness of the proposed statistic is illustrated with an application to experimental data generated to identify genes involved in the response of cultured cells to oncogenic mutations.

email: andrei_yakovlev@urmc.rochester.edu

APPROXIMATING PAIRWISE ALIGNMENT SCORES WITH GENERAL ASSUMPTIONS

Lily Wang*, Vanderbilt University
Pranab K. Sen, University of North Carolina at Chapel Hill

Alignment methods are useful tools for discovering important characteristics of biological sequences since their similarities often suggest likeness in structures, functions and relationships in phylogeny. An essential task is the determination of statistical significance of alignment scores. Since only isolated regions of the sequences will remain similar during evolution, researches in recent years have focused on local alignment scores from the best matching segments of two given sequences. Waterman (1995) gives comprehensive review on this topic for i.i.d. sequences. However, biological sequences rarely follow the i.i.d. model. We studied the asymptotic properties of the pairwise alignment scores under the general assumptions where the matching probabilities are possibly different and the positions are Markovian dependent.

email: lily.wang@vanderbilt.edu

AN IMPROVED MULTIPLE ANALYSIS OF ASSOCIATIONS: A METHOD FOR EVALUATING THE INFLUENCE OF MULTIPLE GENES ON A TRAIT OF INTEREST

Amy D. Anderson*, Bioinformatics Research Center, North Carolina State University

We are interested in evaluating the influence of multiple candidate genes on the onset of temporomandibular disorder. An ideal analysis would involve careful modeling of the various effects associated with each genotype or genotype combination (additive effects, dominance effects, various epistatic effects, etc.), but with a large number of genes this type of analysis becomes infeasible. D.E. Comings et al. (Clin Genet 2000: 57:178--196) proposed an analysis technique that they call Multiple Analysis of Associations, in which they first reduce each gene of interest to a single quantitative variable, then analyze these variables together in a multiple regression. In practice, the method for reducing genotypes to single variables has been based upon whether the gene seems to act in a simple dominant, recessive, or codominant fashion in the data set upon which the analysis is to be run. In our work, we modified this method to (1) allow a more flexible approach to collapsing genotypes within genes into meaningful single variables and (2) incorporate bootstrapping and permutation testing to take into account the fact that the same data set is used for determining the form of the genetic variables and running the multiple regression.

email: amya@statgen.ncsu.edu

COMPARISON OF METHODS FOR THE ANALYSIS OF ALLELIC LOSS DATA

Lei Shen*, Ohio State University

Allelic loss data, binary indicators of loss of heterozygosity at multiple marker positions on one or more chromosomes, have been widely used to identify putative tumor-suppressor genes that play an important role in tumorigenesis. This type of data often show high percentages of missing (noninformative) values and positive correlations between markers on the same chromosome. Methods that have been used to analyze allelic loss data include simple ad hoc methods, a likelihood approach based on a parametric instability- selection model proposed by Newton et al. (1998) and Newton and Lee (2000), a nonparametric method by Miller et al. (2003), changepoint estimation by Albert et al. (2004) and a marginal model approach that we developed and applied to a breast cancer dataset (Fukino et al. 2004). In this talk, we study relationships between various models, assess plausibility of several key assumptions, and compare the performance of different methods.

email: lshen@sph.osu.edu

GENOME PHYLOGENETIC ANALYSIS BASED ON EXTENDED GENE CONTENTS

Xun Gu, Iowa State University
Hongmei Zhang*, University of West Florida

With the rapid growth of entire genome data, whole-genome approaches such as gene content become popular for genome phylogeny inference, including the tree of life. However, the underlying model for genome evolution is unclear, and the proposed (ad hoc) genome distance measure may violate the additivity. In this article, we formulate a stochastic framework for genome evolution, which provides a basis for defining an additive genome distance. However, we show that it is difficult to utilize the typical gene content data i.e., the presence or absence of gene families across genomes to estimate the genome distance. We solve this problem by introducing the concept of extended gene content; that is, the status of a gene family in a given genome could be absence, presence as single copy, or presence as duplicates, any of which can be used to estimate the genome distance and phylogenetic inference. Computer simulation shows that the new tree-making method is efficient, consistent, and fairly robust. The example of 35 microbial complete genomes demonstrates that it is useful not only to study the universal tree of life but also to explore the evolutionary pattern of genomes.

email: hzhang@uwf.edu

LIKELIHOOD FORMULATION OF PARENT-OF-ORIGIN EFFECTS FOR COMPLEX HUMAN DISEASES: CHARACTERIZING IMPRINTED GENES FOR DEVELOPMENTAL DYSLEXIA

Wei Hou*, University of Florida
Cynthia W. Garvan, University of Florida
Jason G. Craggs, University of Florida
George W. Hynd, Purdue University
Rongling Wu, University of Florida

The same allele of a gene may be expressed differently in offspring, depending on its maternal or paternal origin. This phenomenon, referred to as parent-of-origin effect, has been thought to be more common in the developmental control of diseases than previously appreciated. We have developed a likelihood-based method for testing for parent-of-origin effect in complex human diseases. The likelihood formulation models the transmission process of segregating genes from parents to their offspring during meioses, and incorporates the parent-dependent genetic effects of these genes within the framework of finite mixture models. The method implements ascertainments, allowing for differential maternal and paternal ascertainment probabilities. We use this method to estimate major genes associated with parent-of-origin effects of developmental dyslexia in 55 families where at least one sibling was diagnosed with a neuro linguistic processing deficit disorder.

email: whou@biostat.ufl.edu

CHARACTERIZING THE GENETIC STRUCTURE OF POPULATIONS

Xi Chen*, North Carolina State University
Bruce S. Weir, North Carolina State University

The need to characterize the genetic structure of natural populations, especially of humans, has increased with recent large-scale disease association studies. The classical moment estimators for Wright's 'F-statistics' have low bias but large variance. Maximum likelihood estimates based on assumed normality of allele frequencies over populations have proven difficult to obtain in practise. There is hope that the Dirichlet distribution will provide a more robust framework, in spite of the implict assumptions for that distribution not applying to microsatellite markers. The behavior of moment and MLE estimates for data generated by both forward and coalescent simulation will be shown, along with some preliminary estimates for very dense SNP maps.

email: xchen@stat.ncsu.edu

ESTIMATING QTL PARAMETERS UNDER SELECTIVE GENOTYPING

Jaya M. Satagopan*, Memorial Sloan-Kettering Cancer Center
Saunak Sen, University of California, San Francisco
Gary A. Churchill, Jackson Laboratory

Identifying genetic loci associated with quantitative traits is often a multi-step process. Selective genotyping is an outcome-based sampling approach whereby subjects with extreme trait values are genotyped more often than intermediate ones. When subjects are selectively sampled in this manner, a conditional likelihood is a natural way to estimate model parameters. This talk examines the efficiency of selective genotyping by evaluating the properties of the model parameters and the variance estimates derived using a conditional likelihood.

email: satagopj@mskcc.org

45. A PRACTICUM ON MULTISTATE SURVIVAL MODELS

A SHORT SURVEY OF MULTISTATE MODELS

David Oakes*, University of Rochester Medical Center

This introductory talk will address some strategic aspects of multistate modeling. Topics that will be discussed include the choice of time-scale, use of intensity (hazard)- based models as opposed to accelerated life models, marginal vs conditional analysis and the use of frailties to model heterogeneity. Emphasis will be placed on understanding the differing interpetations of parameter estimates in different models.

email: oakes@bst.rochester.edu

PRACTICAL ASPECTS OF MULTI-STATE MODELS

Terry M. Therneau*, Mayo Clinic

The use of a Cox model program, along with careful creation of a data set, is a fairly simple and fast way to approach the analysis of data where subjects can move between multiple states. We will look at some of the practical aspects of such analysis, including creation of the data, choice of a time scale, summarization of results in terms of ``survival'' or hazard curves, and the impact of random effects terms.

email: therneau.terry@mayo.edu

MODELS AND STATISTICAL METHODS FOR THE CURRENT LEUKEMIA FREE SURVIVAL FUNCTION

John P. Klein*, Medical College of Wisconsin

Donor leukocyte infusions (DLI) have become a new therapy for patients who relapse after a bone marrow transplant (BMT). With this therapy patients may be in one of several transient states (alive in remission, alive in first post BMT, alive in second remission following a DLI) or they may be dead. Of interest is inference for the "Current Leukemia Free Survival Function," (CLFS) defined as the probability a patient is alive and in first or second remission post BMT. In this talk we discuss inference for a multistate model which allows us to estimate the CLFS function. We show how summary estimates can be constructed using either a non homogenous Markov process approach or by using an estimator for transient state probabilities based on the difference of Kaplan- Meier estimators. We look at regression models for the CLFS based on a synthesis of regression models for each of the transition intensities or based on a pseudo-value approach proposed by Andersen et al (Biometrika 90, 2003) which allows direct modeling of the CLFS function. The techniques are illustrated on data from the Center for International Blood and Marrow Transplantation.

email: klein@mcw.edu

46. STATISTICS IN DISEASE ECOLOGY

SPATIAL CONTACT NETWORKS AND TIMING OF OUTBREAKS IN EPIDEMIC METAPOPULATIONS: THEORY, DATA, AND STATISTICS

Ottar N. Bjornstad*, Pennsylvania State University

Infectious diseases provide a particularly clear illustration of the spatio-temporal underpinnings of consumer-resource dynamics. The paradigm is the locally unstable, oscillatory dynamics of extremely contagious, directly transmitted, immunizing infections caused by morbilliviruses for which more or less irregular epidemics are interspersed by prolonged periods of local extinction of the parasite. Spatial transmission and 'recolonization' in such systems are ultimately tied to movement by the hosts. The network of spatial spread may therefore be related to the transportation network within the host metapopulation. I discuss two critical issues in the spatiotemporal dynamics of host-pathogen interactions. First, how do rates of movement of hosts between patches affect the timing and predictability of outbreaks? Second, how can we reconstruct topology of transportation networks from outbreak data? I address the first using stochastic epidemic models, and the second the models' associated hazard likelihoods. The theory and methods are discussed with particular reference to measles. I propose a gravity model for the spatial transmission networks and discuss how recurrent epidemics may either be periodic and relatively predictable or erratic and unpredictable depending on the strength of spatial transmission.

email: onb1@psu.edu

INTERVAL TIME LAG MODELS FOR ENVIRONMENTAL EFFECTS ON ARBOVIRUS POSITIVE MOSQUITO POPULATIONS

Frank C. Curriero*, Johns Hopkins University
Scott M. Shone, Johns Hopkins University
Greg E. Glass, Johns Hopkins University

It has long been recognized that arthropod populations fluctuate with changes in environmental conditions and these changes occur at various spatial and temporal scales. Changes in meteorological conditions, such as temperature, precipitation events, relative humidity and wind speed impact both the population size (through changes in survival and reproduction) as well as the ability to sample individual vector populations. Field experience suggests that the duration of these environmental effects on vectors often extends over a range or interval of time rather than a single point in time. In this talk we extend the notion of lagged environmental effects summarized over a time interval to models of arbovirus positive mosquito populations. Daily pools of Culiseta melanura, the predominant mosquito for Eastern Equinine Encephalitis, were trapped at a sites along the Eastern Shore of Maryland and tested for virus. Cross correlation maps are introduced as a graphical tool for visualizing time lagged associations of environmental effects on positive pools of Culiseta melanura. Models based on environmental effects summarized over time intervals reproduce the dynamics in the virus rates better than their counterparts based on lagged effects at single points in time.

email: fcurrier@jhsph.edu

REAL-TIME SPATIAL PREDICTION OF INFECTIOUS DISEASE: EXPERIENCE OF NEW YORK STATE (USA) WITH WEST NILE VIRUS AND FUTURE DIRECTIONS FOR IMPROVED SURVEILLANCE

Glen D. Johnson*, New York State Department of Health

Infectious disease surveillance has become an international top priority due to the perceived risk of bioterrorism. This is driving the improvement of real- time geo-spatial surveillance systems for monitoring disease indicators, which is expected to have many benefits beyond detecting a bioterror event. New York State (USA) has concurrently implemented a working system in response to West Nile Virus, which first appeared in the western hemisphere in New York in 1999 and has subsequently become a pandemic. Since American Crows (Corvus brachyrhynchos) provide a good indicator of viral activity due to their high case fatality rate, this system includes weekly assessments of dead crow clusters using both the binomial spatial scan statistic and kernel density smoothing. A retrospective study of year 2002 data will be presented, showing how spatial clusters of dead crows are significantly associated with human cases, after adjusting for geographic region, time, human population density and age distribution. It will be argued that this system, and infectious disease surveillance in general, can be improved by complementing spatial cluster detection of an outcome variable with predictive "risk mapping" that incorporates spatial- temporal data on the environment, climate and human population through the flexible class of generalized linear mixed models.

email: gdj01@health.state.ny.us

---

47. STATISTICAL METHODS FOR REPRODUCTIVE EPIDEMIOLOGY

OVERVIEW OF METHODOLOGIC CHALLENGES FACING REPRODUCTIVE EPIDEMIOLOGY

Germaine M. Louis*, National Institute of Child Health & Human Development

Reproductive epidemiology focuses on fecundity (biologic capacity for reproduction) and fertility (live births) outcomes. Determinants of each are relatively unexplored. The effect of previous pregnancy outcomes on subsequent pregnancy intentions underscores the interplay between biology and behavior. Virtually no study has followed couples through completed reproductive performance. To minimize bias and maximize efficiency, study design and analysis must be responsive to the clustering of pregnancy outcomes and correlated nature of exposures. Timing of parental exposures at critical phases (e.g., spermatogenesis, ovulation, implantation) is essential for assessing etiologic relations. The dynamic physiologic state of pregnancy impacts accurate exposure assessment for pregnancy dependent outcomes. An overview of human reproduction and development with regard to biologic critical windows and human behavior is presented along with issues such as competing risks, left censoring and changing paternity to identify the opportunities for statisticians to develop and apply methods relevant for reproductive epidemiology.

email: louisg@mail.nih.gov

STATISTICAL METHODS FOR STUDYING GENETIC CONTRIBUTIONS TO BIRTH DEFECTS

Clarice R. Weinberg*, National Institute of Environmental Health Sciences
David M. Umbach, National Institute of Environmental Health Sciences

The etiology of birth defects is "complex," because both genetic and environmental factors contribute to risk. An additional complication for birth defects is that both the fetal and the maternal genotypes can influence risk. Consider a design where one studies only "triads" made up of affected babies and their parents. Under a simplifying assumption of genetic mating symmetry, a loglinear analysis can efficiently detect linkage disequilibrium, with no need to assume Hardy-Weinberg equilibrium in the source population. The relative risks associated with one copy or two copies of a variant allele can also be estimated, both for effects that work through the fetus and for those that work through expression of maternal genes during gestation. No inheritance model needs to be specified (e.g. dominant or recessive) and families with a missing parent can also be included. The efficiency of this design/analysis will be compared with that of a case-control design where mother-child pairs are sampled and the unit of analysis is taken as the pair. The proposed approach has broader application to other phenotypes expressed early in life, e.g., to etiologic studies of childhood cancers.

email: weinber2@niehs.nih.gov

MARGINAL REGRESSION FOR RECURRENT MARKED POINT PROCESS DATA

Patrick J. Heagerty*, University of Washington

Epidemiologic studies investigating the role of specific occupational or lifestyle exposures on birth outcomes need to consider factors that influence the level of exposure and the timing of births. In certain prospective studies multiple births per woman are recorded. In these situations longitudinal regression methods may allow exposure to be correlated with measured characteristics of a pregnancy or birth such as the presence of specific malformations or the birth weight. This talk will outline a general statistical formulation for such reproductive recurrent event data where interest is in the characteristics of the clinical event, or so-called marked point process data. We will show that key assumptions need to be satisfied regarding both the exposure process and the birth process in order for valid use of mixed model methods or use of any non-independence working correlation generalized estimation equation (GEE) approach.

email: heagerty@u.washington.edu

THE DESIGN OF THE NATIONAL CHILDREN'S STUDY: PROBABILITY SAMPLE, MEDICAL CENTER MODEL, OR WHAT?

Roderick J. A. Little*, University of Michigan

The National Children's Study is a massive longitudinal study of determinants of childhood illness. Currently in an extensive planning phase, the study may be the most ambitious and wide-ranging longitudinal observational study ever undertaken. A major issue is the basic design of the study, and in particular whether it can and should be based on a probability sample of the population, or should follow the medical-center based model adopted by a number of previous epidemiologic studies in this area. Views on this controversial issue range from those who believe a non-probabilistic design is essentially unscientific, to others who believe a probabilistic design is impractical and unnecessary, and sacrifices internal for external validity. In the talk I will discuss some of the arguments presented for the various approaches, including hybrid designs. I give my personal views on the issue, and discuss the current plans for the design of the study.

email: rlittle@umich.edu

48. MODELING BRAIN IMAGES—THE EFFECTS OF SPACE, TIME, AND INDIVIDUALITY

REVEALING BRAIN ACTIVITY WITH FILTERS

Kary L. Myers*, Carnegie Mellon University

In optical imaging experiments, scientists use a digital video camera to record changes on the surface of the brain. The technique offers tremendous spatial and temporal resolution for functional neuroimaging: In the experiments I study, each pixel is about 150 square microns (the size of a very small cell), and the video arrives at a rate of 30 frames per second. Unfortunately, the data typically have signal-to-noise ratios close to 0.001, making the recorded brain activity difficult to map. In this talk I'll discuss the use of Kalman filtering to reveal this brain activity. Several features of the filter approach make it a good candidate for use with optical imaging data: Brain activity can be modeled as a latent variable to be estimated; the model can incorporate additional inputs to the system like physiological changes and experimental stimuli while using significantly fewer parameters than a comparable regression model; and the filter can operate in an online fashion as part of the data acquisition system.

email: kary@stat.cmu.edu

SPATIO-TEMPORAL MODELING OF LOCALIZED BRAIN ACTIVITY

DuBois Bowman*, Emory University

Functional neuroimaging, including PET and fMRI, plays an important role in identifying brain regions associated with experimental stimuli and perhaps psychiatric disorders. PET and fMRI produce massive data sets that contain temporal correlations from repeated scans and complex spatial correlations. Several methods exist for handling temporal correlations. Despite the presence of spatial correlations between image voxels, conventional methods perform voxel-by-voxel analyses of measured brain activity. We propose a two-stage spatiotemporal model for estimation and testing of localized activity. Our second- stage model specifies a spatial autoregression, capturing correlations within neural processing clusters defined by a data-driven cluster analysis. We use maximum likelihood methods to estimate parameters of our spatial model. Our approach protects against type-I errors, detects localized and regional activations, provides information on functional connectivity in the brain, and establishes a framework to produce individualized spatially smoothed activity maps. We illustrate our approach using PET data from a study of working memory in individuals with schizophrenia.

email: dbowma3@sph.emory.edu

ROBUST AND LOCAL NONSPHERICITY MODELING FOR SECOND-LEVEL PET AND FMRI ANALYSIS

Thomas E. Nichols*, University of Michigan
Jeanette Mumford, University of Michigan
Wen-Lin Luo, University of Michigan

Users of fMRI & PET data have recently discovered the importance of mixed effects models for inference with group (multisubject) data. When an effect can be summarized with one image per subject (say a mean difference image, or a contrast image from a GLM), there are straightforward summary-statistic methods available. When there are multiple effects of interest per subject, nonsphericity must be accounted for. In this context, nonsphericity refers to heterogeneous variance over effects and intrasubject correlation. The widely used SPM software uses iterative methods to find a nonsphericity estimate which is pooled over the whole brain. Instead, we propose the use of Generalized Estimating Equation (GEE) methods for non- iterative yet local covariance estimation. Critically, our GEE method provides consistent variance estimates, while SPM must assume spatially homogeneous nonsphericity for validity. We evaluate our method with simulations and real data. From simulations we find that GEE-based variance estimates are less biased than SPM's, and from real data we find evidence that intrasubject correlation varies substantially over the brain. These results suggest that local co-variance estimation is just as important as local variance estimation.

email: nichols@umich.edu

49. CROSSOVER DESIGNS FOR THE PHARMACEUTICAL INDUSTRY

OPTIMAL AND EFFICIENT CROSSOVER DESIGNS WHEN SUBJECT EFFECTS ARE RANDOM

John Stufken*, University of Georgia

Crossover designs are often evaluated under the assumption that subject effects are fixed. One justification for this is that most information about treatment comparisons is based on within subject information. But how efficient are designs that are optimal for fixed subject effects when the subject effects are really random? Which designs are optimal when subject effects are really random, and how does this change with the size of the subject effects variance relative to the size of the random error variance? We investigate these questions in the presence of carry- over effects for the situation that the number of periods is at most equal to the number of treatments.

email: jstufken@stat.uga.edu

EFFICIENT DESIGNS FOR EXPERIMENTS WITH MULTIPLE TREATMENTS PER SUBJECT

James L. Rosenberger*, Pennsylvania State University

Crossover designs arise in many situations where subjects are used for multiple treatments. This talk considers situations where there are more treatments than observational periods and subjects are reused in multiple sessions. Repeated use of subjects in multiple sessions allows for control of both subject effects and order and carryover effects. An example is described with 32 cocoa treatments tested by taste panelists in sessions which could measure 4 treatments in sequence. With 32 subjects available, the study was designed to allow testing for residual carryover effects of first, second and third order. However, subjects were not balanced with respect to gender and flavor taster status, which provided an additional challenge in modeling the results. Response functions were used to model the effect of the composition of the cocoa.

email: JLR@psu.edu

CROSSOVER DESIGNS FOR COMPARING TEST TREATMENTS WITH A CONTROL

Sam Hedayat*, University of Illinois at Chicago

Although comparing t test treatments with one control treatment in a crossover design is an important issue, little research has addressed this issue when the number of periods is less than number of treatments t+1. This talk will present the latest results for this topic. We will give sufficient conditions for a crossover design to be simultaneously A-optimal and MV-optimal in a very large and appealing class of crossover designs.Some optimal/efficient designs are constructed for some practical parameters. The robustness of these newly discovered crossover designs under different models will be discussed.

email: hedayat@uic.edu

ON OPTIMAL CROSS-OVER DESIGNS WHEN CARRY-OVER EFFECTS ARE PROPORTIONAL TO DIRECT EFFECTS

R. A. Bailey, Queen Mary, University of London
J. Kunert*, Universität Dortmund

There are a number of different models for cross-over designs which take account of carry-over effects. Since it seems plausible that a treatment with a large direct effect should generally have a larger carry-over effect, Kempton, Ferris and David (2001) considered a model where the carry-over effects are proportional to the direct effects. The advantage of this model lies in the fact that there are fewer parameters to be estimated. Its problem lies in the non-linearity of the estimates. Kempton, Ferris and David (2001) considered the least squares estimate. They point out that this estimate is asymptotically equivalent to the estimate in the linear model which assumes the true parameters to be known. For this estimate they determine optimal designs numerically for some cases. The present paper generalizes some of their results. Our results are derived with the help of a generalization of the methods used in Kunert and Martin (2000).

email: joachim.kunert@udo.edu

THE PREDICTIVE DISTRIBUTION AND DIAGNOSTIC ACCURACY

Lyle D. Broemeling*, University of Texas M. D. Anderson Cancer Center
Marcella Johnson, University of Texas M. D. Anderson Cancer Center

Using the Bayesian predictive distribution of future observations, the area under the ROC curve is computed. It is assumed that the diagnostic marker is ordinal, then using a conjugate Dirichlet prior for the parameters of the multinomial distribution, the posterior distribution of the ROC area is determined. The methodology is illustrated with multi-reader, multi-modality scenarios taken from a diagnostic imaging setting.

email: lbroemel@mdanderson.org

PEAK-PICKING ALGORITHM FOR LC-ESI-FT MASS SPECTROMETRY DATA

Jeanette E. Eckel-Passow*, Mayo Clinic College of Medicine
Ann L. Oberg, Mayo Clinic College of Medicine
Terry M. Therneau, Mayo Clinic College of Medicine
Chris J. Mason, Mayo Clinic College of Medicine
David C. Muddiman, Mayo Clinic College of Medicine

Liquid-chromatography electrospray-ionization Fourier- transform mass spectrometry (LC-ESI-FT-MS) is a potentially superior biomarker discovery platform because it offers outstanding mass accuracy, mass precision and resolving power over a broad m/z range. These data are extremely complex such that a single sample produces readings at thousands of m/z values over a wide dynamic range and across an entire elution window. The first step in analyzing such data is to reduce the complexity of each spectra via a peak-picking algorithm that is capable of (1) identifying an isotope cluster, where a cluster represents a series of peaks that are one neutron apart in weight, (2) determining the charge state of the cluster and subsequently fitting the cluster to determine the neutral monoisotopic mass and (3) determining a value for the abundance of the cluster. The proposed peak-picking algorithm is demonstrated using spike-in data where nine oligomers with known m/z values were mixed into a parent sample and ten analysis samples were prepared from the parent, each with a different amount of analyte. A sensitivity analysis of the required parameters in the fitting routine is performed, specifically monoisotopic mass, expected number of extra neutrons and peak spacing.

email: eckel@mayo.edu

A NOVEL ALGORITHM FOR MALDI-TOF MS DATA PROCESSING USING MATHEMATICAL TOOLS

Shuo Chen*, East Tennessee State University
Don Hong and Yu Shyr, Vanderbilt University

Mass Spectrometry, especially matrix assisted laser desorption/ionization (MALDI) time of flight (TOF), is emerging as a leading technique in the proteomics revolution. It can be used to find disease-related protein patterns in mixtures of proteins derived from easily obtained samples. In this paper, a novel algorithm for MALDI-TOF MS data processing is developed. The software design includes the application of splines for data smoothing and baseline correction, wavelets for adaptive denoising, multivariable statistics techniques such as clustering analysis, and signal processing techniques to evaluate the complicated biological signals. A MatLab implementation shows the processing steps consecutively including step-interval unification, adaptive wavelet denoising, baseline correction, normalization, and peak detection and alignment for biomarkers discovery.

email: hitcx@hotmail.com

RESOLVING THE DEGREES OF FREEDOM ISSUE CONCERNING THE DORFMAN-BERBAUM-METZ AND OBUCHOWSKI-ROCKETTE METHODS FOR RECEIVER OPERATING CHARACTERISTIC (ROC) DATA

Stephen L. Hillis*, Iowa City V.A. Medical Center

Several methods are currently used for analyzing multi- reader ROC studies. The Dorfman-Berbaum-Metz (DBM) method consists of a conventional mixed model analysis applied to pseudovalues of ROC parameters computed by jackknifing cases separately for each reader-modality combination. The Obuchowski-Rockette (OR) method consists of a mixed model analysis of the reader-modality ROC parameter estimates, with the analysis adjusted to correct for the correlations between and within readers. Recently it has been shown that when DBM and OR utilize the same accuracy measure and covariance estimation method, the test statistics will be equal but inferences can vary considerably if different denominator degrees of freedom methods are used; this clearly is an unacceptable situation. To resolve this issue, we show that the original OR denominator degrees of freedom method is not appropriate and propose a revised method for estimating the denominator degrees of freedom. In simulations we show that the revised method performs similarly to the DBM method, with the revised OR method tending to be slightly more liberal than DBM. This is to be expected, since OR treats error covariances as known while DBM treats them as unknown.

email: steve-hillis@uiowa.edu

A FRAMEWORK FOR THE STUDY OF THE PREDICTIVE ACCURACY OF DIAGNOSTIC TESTS

Shang-Ying Shiu*, Brown University
Constantine Gatsonis, Brown University

In this paper, we examine the effect of the threshold for test positivity on the positive and negative predictive values of diagnostic tests. We define the Predictive Receiver Operating Characteristic Curve (PROC), which summarizes the pairs of positive and negative predictive values as the threshold for test positivity varies. We study the curve's geometric properties, and present methods to quantify and compare predictive performance of tests using this PROC curve.

email: shiu@stat.brown.edu

A PERMUTATION TEST SENSITIVE TO DIFFERENCES IN AREAS FOR COMPARING ROC CURVES FROM A PAIRED DESIGN

Andriy I. Bandos*, University of Pittsburgh
Howard E. Rockette, University of Pittsburgh
David Gur, University of Pittsburgh

The Area Under the ROC Curve (AUC) is a widely accepted summary index of the overall performance of diagnostic procedures and the difference between AUCs is often used when comparing two diagnostic systems. We propose the permutation test that uses non-parametric estimate of the AUC difference and being formally a test for equality of ROC curves is sensitive primary to the alternatives of AUC difference. The operating characteristics of the proposed test were evaluated using extensive simulations over a wide range of parameters. The proposed procedure can be easily implemented in experimental ROC datasets. For small samples and for underlying parameters that are common in experimental studies in diagnostic imaging the test possesses good operating characteristics and is more powerful than the conventional non-parametric procedure for AUC comparisons. We also derived an asymptotic version of the test which uses an exact estimate of the variance in the permutation space and provides a good approximation even when the sample sizes are small. This asymptotic procedure is a simple and precise approximation to the exact test and is useful for large sample sizes where the exact test may be computationally burdensome.

email: anb61@pitt.edu

LINEAR MIXED MODELS WITH SKEWED T DISTRIBUTIONS

Tianyue Zhou*, University of Illinois at Urbana-Champaign

Linear mixed models based on normality assumptions are widely used in health related studies. Although the normality assumption leads to simple, mathematically tractable, and powerful tests, violation of the normality assumption may invalidate statistical inference. Transformation of variables is sometimes used to make normality approximately true. In this paper we introduce another approach by replacing the normal distributions in linear mixed models by skew t distributions, which allow skewness and heavy tails. Likelihood-based inference is developed, followed by illustrations using simulated data and real data. Our analysis on some temporal variables in swallowing studies shows that skew t errors are often realistic even on log-transformed response variables.

email: tzhou1@uiuc.edu

ANALYTICAL METHODS FOR COMPLIANCE WITH LONGITUDINAL ASSESSMENTS IN CLINICAL STUDIES

Stephanie R. Land*, University of Pittsburgh
Ritter Marcie, University of Pittsburgh

In longitudinal clinical studies, tremendous resources are often required to assure continued compliance with study assessments. This presentation is motivated by particular examples of behavioral and health outcomes (such as quality of life) studies in clinical trials, in which participants are asked to complete questionnaires at fixed intervals over several months or years. Efforts to improve compliance may benefit from a greater understanding of the factors that influence compliance. In particular, interventions designed to increase compliance can be tested in a randomized design. In this presentation I will describe such a randomized study being conducted at my institution, illustrate a mixed model approach to estimate the effects of candidate factors, and describe a simulation method to determine sample size requirements for such a study.

email: stephanie.land2@verizon.net

A DIAGNOSTIC TEST FOR THE MIXING DISTRIBUTION IN A GENERALIZED LINEAR MIXED MODEL

Eric J TChetgen*, Harvard University
Brent Coull, Harvard University

We introduce a diagnostic test for the mixing distribution in a generalised linear mixed model. The test is based on the difference between the marginal maximum likelihood and conditional maximum likelihood estimates of a subset of the fixed effects in the model. We derive the asymptotic variance of this difference, and propose a test statistic that has a limiting chi-square distribution under the null hypothesis that the mixing distribution is correctly specified. For the important special case of the logistic regression model with random intercepts, we evaluate via simulation the power of the test in finite samples under several alternative distributional forms for the mixing distribution. We illustrate the method by applying it to data from a clinical trial investigating the effects of hormonal contraceptives in women.

email: etchetge@hsph.harvard.edu

BAYESIAN COVARIANCE SELECTION IN GENERALIZED LINEAR MIXED MODELS

Bo Cai*, National Institute of Environmental Health Sciences
David B. Dunson, National Institute of Environmental Health Sciences

The generalized linear mixed model (GLMM), which extends the generalized linear model (GLM) to incorporate random effects characterizing heterogeneity among subjects, is widely used in analyzing correlated and longitudinal data. Although there is often interest in identifying the subset of predictors that have random effects, random effects selection can be challenging, particularly when outcome distributions are non-normal. This article proposes a fully Bayesian approach to the problem of simultaneous selection of fixed and random effects in GLMMs. Integrating out the random effects induces a covariance structure on the multivariate outcome data, and an important problem which we also consider is that of covariance selection. Our approach relies on variable selection-type mixture priors for the components in a special LDU decomposition of the random effects covariance. A stochastic search MCMC algorithm is developed, which relies on Gibbs sampling, with Taylor series expansions used to approximate intractable integrals. Simulated data examples are presented for different exponential family distributions, and the approach is applied to discrete survival data from a time- to-pregnancy study.

email: cai@niehs.nih.gov

SMALL SAMPLE INFERENCE FOR CLUSTER SAMPLES WITH GAUSSIAN DATA

Jacqueline L. Johnson*, University of North Carolina at Chapel Hill
Diane J. Catellier, University of North Carolina at Chapel Hill
Keith E. Muller, University of North Carolina at Chapel Hill

Linear mixed models provide a popular framework for analyzing Gaussian data from cluster samples where the number of observations in each cluster varies. The methods perform well in large samples, but can give very optimistic inference in small samples, including greatly inflated test size. If all clusters are the same size (complete balanced data), Muller, Coffey, and Gurka (in review) showed how to transform a mixed model with compound symmetric covariance into a set of univariate linear models, which provide optimal exact (small-sample) inference for all main effects and interactions of the original model. We extend the result to models with unequally sized clusters (missing data), and provide exact inference for within cluster and within by between cluster effects, and accurate approximations for between cluster effects. Computing the estimates does not require any iteration.

email: jjohnson@bios.unc.edu

MODELING INTER-RATER AGREEMENT USING MIXED MODELS

Kerrie Nelson*, University of South Carolina
Don Edwards, University of South Carolina

Medical diagnoses are often based on the opinions of physicians and the agreement between them. Current methods for assessing inter-rater agreement, including Kappa, can be prone to bias or do not easily incorporate multiple raters or unbalanced data. The use of a crossed generalized linear mixed model provides a flexible approach to yield inference about inter-rater agreement in a general underlying diagnostic process. The models are applied to cancer data and compared with currently used methods.

email: kerrie@stat.sc.edu

A NON-LINEAR MIXED EFFECT MODEL FOR HEPATITIS C VIRAL DYNAMICS

Abdus S. Wahed*, University of Pittsburgh
Kyungah Im, University of Pittsburgh
Thelma Wiley, Rush University
Steven Belle, University of Pittsburgh

Treatment of patients infected with Hepatitis C Virus (HCV) using antiviral therapy such as Peginterferon or Peginterferon in combination with Ribavirin has been proven to be effective. The efficacy of these antiviral agents is often assessed by the viral levels during the treatment period and by the post-treatment sustained virological response. Compartment models such as the one described by Neumann et al. (Science, 1998, vol 282, pp 103-107) are routinely used for characterizing the HCV dynamics. The Neumann et al. model fails to take into account the pharmacokinetic behavior of the drug (especially, the interferon level in the blood serum over time). Also, the model does not explicitly include other patient characteristics such as age, race, or body mass index that might explain the variation in the viral level. We extend the Neumann et al. model to incorporate the pharmacokinetic behavior of Peginterferon and other patient characteristics on HCV dynamics. Efficacy of the drug, virus clearance rate and the effect of patient characteristics on viral kinetics will be estimated. Large sample behavior of the proposed model and the properties of the estimated parameters will be investigated using simulated data. We will demonstrate an application using data from VIRAHEPC clinical trial.

email: wahed@pitt.edu

---

52. BAYESIAN METHODS

BAYESIAN INFERENCE IN GENERALIZED ADDITIVE MIXED MODELS WITH NONPARAMETRIC RANDOM EFFECTS

Yisheng Li*, University of Texas M. D. Anderson Cancer Center
Xihong Lin, University of Michigan

We propose a nonparametric Bayesian generalized additive mixed model for correlated data. The model allows for additive functional dependence of an outcome variable on covariates by using nonparametric regression and accounts for correlation between observations using random effects. Partially improper integrated Wiener priors are used for the nonparametric functions and the resulting estimators are cubic smoothing splines. The distribution of the random effects is assumed nonparametric with a Dirichlet process prior. Systematic inference on model parameters can be made within a modified nonparametric generalized linear mixed model framework. Computation is carried out using Gibbs sampling. We illustrate the proposed approach by analyzing an infectious disease data set and evaluate its frequentist properties through simulations.

email: ysli@mdanderson.org

HIERARCHICAL MODELS FOR A TIME SERIES ON MARIJUANA ABUSE AMONG HOSPITAL EMERGENCY ROOM ADMISSIONS

Li Zhu*, Texas A&M University
Dennis Gorman, Texas A&M University
Scott Horel, Texas A&M University
Wen Tan, Texas A&M University

Starting in late 1990s, electors in eight states have voted in favor of propositions that legalize the medical use of marijuana. The federal government opposes the introduction of medical marijuana laws on a number of grounds. In this study we test the hypothesis that the introduction of medical marijuana laws is followed by an increase in the use of the drug among a high risk group – hospital emergency room admissions. We use data from Drug Abuse Warning Network (DAWN) program to assess pre-law and post-law trends in marijuana and other drug use. Both log- linear and Poisson regression models are fitted. Bayesian hierarchical structure is built to model the dependence of drug abuse on time, place, medical marijuana law status, and demographic variables. The results suggest that the cities in states that passed medical marijuana law experienced higher increase in drug abuse.

email: lizhu@srph.tamhsc.edu

BAYESIAN ADAPTIVE REGRESSION SPLINES FOR HIERARCHICAL DATA

Jamie L. Bigelow*, National Institute of Environmental Health Sciences
David B. Dunson, National Institute of Environmental Health Sciences

Motivated by studies of reproductive hormone profiles in the menstrual cycle, this article considers methodology for hierarchical functional data analysis. Methods are needed that avoid standardizing cycle lengths, instead allowing for flexible relationships between the hormone level and covariates, including timing relative to reference points, such as the start of the cycle and ovulation. In addition, it is necessary to account for within- woman dependency in the hormone trajectories from multiple cycles. We propose a Bayesian approach based on a hierarchical generalization of the Bayesian multivariate adaptive regression splines approach of Holmes and Mallick (2001). Our formulation allows for an unknown set of basis functions characterizing the overall covariate effects and woman-specific deviations. A reversible jump Markov chain Monte Carlo algorithm is developed for posterior computation. Applying the methods to data from the North Carolina Early Pregnancy Study, we investigate differences in progesterone profiles between conception and non-conception cycles.

email: jbigelow@bios.unc.edu

BAYESIAN FREE KNOT CURVE FITTING WITH APPLICATIONS TO SPECTRAL DENSITY ESTIMATION

Carsten H. Botts*, University of Florida and Iowa State University
Michael Daniels, University of Florida

We model functional data from many subjects with a smoothing spline linear mixed model. With this mixed model, the expected value for any subject can be written as the sum of a population curve and a subject specific deviate from this population curve. The population curve and the subject specific deviates are both modeled as smooth splines with a total of k and k' knots, respectively. These knots are located at t_k and t_k'. We make inference on the population and subject specific curves by conducting a fully Bayesian analysis on the number of knots corresponding to each curve. To be more specific, we try to sample from the posterior p(k, t_k, k', t_k'|y) using reversible jump MCMC methods. This procedure is complicated by the fact that no analytical form exists for this posterior. We explore a variety of approximations to this likelihood and study how each approximation penalizes mixed linear models with too many knots. We then apply these methods to spectral density estimation in longitudinal studies.

email: cbotts1@cox.net

PREDICTION OF PROTEIN INTER-DOMAIN LINKER REGIONS BY A HIDDEN MARKOV MODEL

Kyounghwa Bae*, Texas A&M University
Christine G. Elsik, Texas A&M University
Bani K. Mallick, Texas A&M University

We wish to predict the inter-domain linker regions in a protein sequence using primary sequence alone, without the requirement of known homology. Identifying linker regions will lead to delineation of domain boundaries. We develop a hidden Markov model (HMM) to model linker/non-linker sequence regions in a protein sequence by exploiting differences in the composition of amino acids between the two regions. We recognize the protein sequence data as continuous data instead of categorical data by using the linker index, which incorporates differences in the composition of amino acids in different regions in protein sequences. We employ an efficient Bayesian estimation of the model through Markov Chain Monte Carlo (MCMC) methods, particularly Gibbs sampling to simulate the parameters from the posteriors. We applied our method to a dataset of protein sequences with domains and inter-domain linkers delineated using the Pfam-A database. The prediction results are superior to simpler methods. Importantly, our method determines the probability that each residue belongs to a linker region, which will lead to further insight into properties of inter-domain linkers.

email: kbae@stat.tamu.edu

REPARAMETERIZATION TO IMPROVE BAYESIAN COMPUTING FOR THE CAR MODEL WITH TWO TYPES OF NEIGHBOR RELATIONS

Yi He*, University of Minnesota
James S. Hodges, University of Minnesota

Recent developments in Bayesian computing allow accurate estimation of integrals, making advanced Bayesian analysis feasible. However, some problems are still difficult, including posterior distributions for variances in hierarchical models. This paper focuses on a particular model, the conditional autoregressive model with two kinds of neighbor relations (2NRCAR). Because of interactions between the two types of neighbor relations, the posterior distributions of the smoothing parameters can take a great variety of shapes, including multiple modes and L-shaped contours, which makes computation challenging. Thus, 2NRCAR makes a nice testbed for MCMC methods for multiple- variance models. This talk compares several reparameterizations of the variance structures and corresponding MCMC samplers. Preliminary analyses show the Besag-Higdon (JRSSB 1999) parameterization with the slice sampler gives faster convergence and better mixing compared to the alternatives. Proposed samplers are applied to periodontal datasets analyzed using the 2NRCAR model.

email: yihe@biostat.umn.edu

CLUSTERING ANALYSIS OF ORDINAL DATA

Xian Zhou*, University of Texas M. D. Anderson Cancer Center
Peter Mueller, University of Texas M. D. Anderson Cancer Center
Nebiyou Bekele, University of Texas M. D. Anderson Cancer Center

Clustering is an important technique in many disciplines such as gene selection, medical image analysis, and market segmentation, etc. Hierarchical and nonhierarchical clustering methods are two broad classes used in the clustering analysis. However, little study has been done on the ordinal data, especially from a Bayesian perspective. Many publications have illustrated the Bayesian clustering approach via mixture models. Motivated mainly by a problem of clustering cancer patients based on their different biomarkers' levels, we propose a Bayesian probabilistic model in estimating the clustering membership using a mixture of Gaussian distributions to tackle the problem of clustering. Reversible jump Markov chain Monte Carlo algorithm is used to identify the total number of clusters. Results are compared with those obtained from K- means method.

email: xianz@mdanderson.org

53. MISSING DATA METHODS

CHALLENGES OF NON-IGNORABLE MISSING DATA IN CLINICAL TRIALS: A PATTERN MIXTURE MODEL APPROACH

G. K. Balasubramani*, University of Pittsburgh
Stephen R. Wisniewski, University of Pittsburgh
James Luther, University of Pittsburgh

A common problem in clinical studies is that data are missing due to dropouts or missed visits. When patients exit the study due to worsening of the depression symptoms or remission, then the missingness is considered to be missing not at random. In most of the studies in depression disorder, the outcome of interest is the 17- item Hamilton Rating scale for depression. There are other measures of depression that are highly correlated with this outcome. This paper analyzes and estimates the non- ignorable missing outcomes. This type of data imputation is challenging due to non-ignorable missing responses in addition to related covariates. We propose a selection model for estimating parameters of a variable known to be highly correlated with the outcome measure and covariates. We build the imputation model with non-missing outcomes and the predicted values from the selection model. The missing non-ignorable values are estimated with the consistent estimates from the imputation model, so that the likelihood is maximized using multiple imputation. The parameters are estimated via Monte Carlo simulation and we evaluate the performance of the maximum likelihood estimates. A pattern-mixture model approach is used to combine the values in selection model and multiple imputation model for estimating parameters.

email: balagk@edc.pitt.edu

REGRESSION ANALYSES WITH DATA MISSING AT RANDOM—AN EXTENSION OF THE EM ALGORITHM

Yang Y. Zhao*, University of Waterloo
Jerald F. Lawless, University of Waterloo
Donald L. McLeish, University of Waterloo

In regression analysis some covariates and/or response data may be missing either by design or happenstance. However, sometimes auxiliary variables for the covariates and/or response are always observable. We will talk about an extension of an EM algorithm which estimates the regression parameters from a semiparametric maximum likelihood. To retain the relative simplicity of the EM algorithm we have to discretize some continuous variables. Some investigation on the consequences of doing this is reported.

email: y3zhao@uwaterloo.ca

A HIERARCHICAL TECHNIQUE FOR ESTIMATING LOCATION PARAMETER IN THE PRESENCE OF MISSING DATA

Sergey S. Tarima*, University of Kentucky
Yuriy G. Dmitriev, Tomsk State University, Russia
Richard J. Kryscio, University of Kentucky

This paper proposes a hierarchical method for estimating the location parameters of a multivariate vector in the presence of missing data. At i^th step of this procedure an estimate of the location parameters for non-missing components of the vector is based on combining the information in the subset of observations with the non- missing components with updated estimates of the location parameters from all subsets with even more missing components in an iterative fashion. If the variance- covariance matrix is known, then the resulting estimator is unbiased with the smallest variance provided missing data are ignorable. It is also shown that the resulting estimator based on consistent estimators of variance- covariance matrices obtains unbiasedness and the smallest variance asymptotically. This approach can also be extended to some cases of non-ignorable missing data. Applying the methodology to a data with random dropouts yields the well- known Kaplan-Meier estimator.

email: stari@ms.uky.edu

EFFECTS OF METHODS OF ESTIMATION OF MISSING DATA

J. Lynn Palmer*, University of Texas M. D. Anderson Cancer Center

Different methods of estimating missing data may result in different conclusions. This study shows results of methods of estimating missing data in a palliative care setting. Due to the fragile health states of some subjects receiving palliative care (treatment that may not be based on increasing survival, but rather on alleviating symptoms), this type of data can be prone to include information that is not missing at random, or that is related to the outcome that is being studied. This report includes results of using three simple methods of imputing data that are currently being used in the palliative care literature, and other methods that take into account variability due to estimation. It emphasizes the importance that should be given to clearly specifying the assumptions behind the method used, since each method may be based upon different assumptions and could result in conflicting conclusions.

email: jlp@odin.mdacc.tmc.edu

CORRELATING TWO CONTINUOUS VARIABLES SUBJECT TO DETECTION LIMITS IN THE CONTEXT OF MIXTURE DISTRIBUTIONS

Haitao Chu*, Johns Hopkins University
Lawrence H. Moulton, Johns Hopkins University
Wendy J. Mack, University of Southern California
Douglas J. Passaro, University of Illinois at Chicago
Paulo F. Barroso, Hospital Universitário Clementino Fraga Filho School of Medicine
Alvaro Muñoz, Johns Hopkins University

In individuals infected with human immunodeficiency virus (HIV), distributions of quantitative HIV RNA viral load measurements may be highly left-censored with an extra spike below the detection limit (LD) of the assay. When estimating the correlation coefficient between two different measures of viral load obtained from each of a sample of patients, a bivariate Gaussian mixture model is recommended to better model the extra spike on [0, LD1] and [0, LD2] when the proportion below LD is incongruent with the left tail of a bivariate Gaussian distribution. Maximum likelihood is used to estimate the parameters of the lower and higher components. To evaluate whether there exists a lower component, we apply a Monte Carlo approach to assess the p-value of the likelihood ratio test and two information criteria. We provide simulation results to evaluate the performance and compare it with two ad hoc estimators and a single component bivariate Gaussian likelihood estimator. These methods are applied to the data from a cohort study of HIV-infected men in Rio de Janeiro, Brazil and the data from the Women's Interagency HIV oral study. These results emphasize the need for caution when estimating correlation coefficients from data with a large proportion of non-detectable values when the proportion below LD is incongruent with the left tail of a bivariate Gaussian distribution.

email: hchu@jhsph.edu

SAMPLE SIZE CALCULATION FOR LONGITUDINAL DATA UNDER MISSING AT RANDOM (MAR)

Susan Halabi*, Duke University Medical Center
Daohai Yu, Duke University Medical Center
Sin-Ho Jung, Duke University Medical Center

In many clinical trials, often a repeated measurement design is used in which individuals are randomly assigned to treatment arms and followed-up over time. Sample size computation is an important component at the design stage. Some patients, however, may miss their clinic visits and the underlying missing data mechanism usually is not missing completely at random (MCAR). Liu and Liang (1997) proposed a method to calculate the sample size and power for correlated observations using the generalized estimating equation (GEE) approach. This method and that of Rochon (1998) do not take into account the mechanism of missing. Jung and Ahn (2003) proposed a sample size formula assuming MCAR. Using the GEE method, we propose closed-form sample size and power formulae for comparing the slopes between two treatment arms of longitudinal data under MAR. Simulations are performed to evaluate the empirical power assuming an auto-regressive or exchangeable correlation structure. This method is illustrated in the design of a clinical study where prostate cancer patients are randomized with equal probabilities to either a placebo or treatment arm. The primary endpoint is the prostate specific-antigen (PSA) slope. PSA measurements are to be collected at baseline, 3, 6, 9 and at 12-months from randomization.

email: susan.halabi@duke.edu

MULTIVARIATE ANALYSIS OF BINARY DATA FROM DRUG SAFETY TRIALS

Bernhard Klingenberg*, Williams College
Alan Agresti, University of Florida

This talk explores global tests of equality between two vectors of binomial probabilities, based on data from two independent, multivariate binary samples. Such data commonly arise in safety studies for newly developed drugs, where the occurrence of several, possibly correlated, side effects are compared under a placebo and a drug treatment. Equality between the treatments is defined as either simultaneous homogeneity in the marginal distributions or as identity of the joint distributions. Following a multivariate approach to inference rather than combining results from univariate tests, we construct binary data analogs of the well known Hotelling T-squared test for multivariate normal data. Likelihood ratio tests based on fitting marginal models for binary data can be computationally intensive for even a moderate number of side effects. Instead, we focus on simpler quadratic form statistics that reduce to well known Wald and score type tests in the univariate case. For either type of statistic, asymptotic inference is often inappropriate due to sparseness, and we also present exact permutation tests. The methods are illustrated with safety data from a Phase II clinical trial.

email: bklingen@williams.edu

A MARGINALIZED DIFFUSION MODEL FOR COMBINING STATE AND NATIONAL LEVEL SURVEY DATA

Diana L. Miglioretti*, Group Health Cooperative
Elizabeth Brown, University of Washington

We propose a marginalized regression approach for combining state (BRFSS) and national (NHIS) level survey data to estimate colorectal cancer screening trends. Interest is in the cumulative incidence of screening in the United States, which is modeled using a mixed-influence diffusion of innovations model. The NHIS data is directly modeled using this national-level model. The BRFSS data is modeled using a state-level (conditional) model to account for clustering within states and missing data. This state-level model is linked to the national model for joint estimation.

email: miglioretti.d@ghc.org

MODELING DIFFERENTIATED TREATMENT EFFECTS FOR MULTIPLE OUTCOMES DATA

Hongfei Guo*, Johns Hopkins University

Multiple outcomes data are commonly used to characterize treatment effects in medical research, for instance, multiple symptoms to characterize potential remission of a psychiatric disorder. Often either a global treatment effect, i.e. assuming a common effect, or individual treatment effects whose magnitude varies by outcomes, are evaluated in the literature. Potentially the global treatment effect may overgeneralize the effect for all the outcomes; however individual treatment effects are complicated to interpret and may suffer power loss by assigning a different effect for each outcome. A better way to summarize the treatment effect may be through patterns of the treatment effects, i.e. ``differentiated effects''. In this paper I propose a two-category model to differentiate treatment effects into two groups. A model fitting algorithm and simulation study are presented, and several methods are developed to analyze heterogeneity presenting in the treatment effects. The method is illustrated using an analysis of schizophrenia symptom data.

email: hfguo@jhsph.edu

AVOIDING TEST SIZE BIAS DUE TO INTERNAL PILOTS WITH GAUSSIAN REPEATED MEASURES

Meagan E. Clement*, University of North Carolina at Chapel Hill
Christopher S. Coffey, University of Alabama at Birmingham
Keith E. Muller, University of North Carolina at Chapel Hill

Uncertainty about the error covariance matrix usually presents the biggest barrier to accurate power analysis in the "univariate" approach to repeated measures analysis of variance. When the planning covariance matrix differs from the true value, Coffey and Muller (Stat Med, 2003) showed that an internal pilot design can be used to maintain expected power or reduce expected sample size. However, they advised against the widespread use of internal pilots in such settings because the unadjusted Geisser-Greenhouse test, based on the total sample covariance estimate, can inflate test size. An unadjusted Box (conservative) test provides one solution to this problem but is extremely conservative. New analytic results provide better solutions by extending univariate results which control bias in test size introduced by an internal pilot. Each method uses a weighted average of the independent covariances estimated in the internal pilot and from information in the total sample which is orthogonal to the internal pilot. For many settings, the methods provide the advantages of internal pilots for repeated measures while avoiding test size bias, even in small samples.

email: clement@email.unc.edu

COMPARING TREATMENT MEANS IN A REPEATED MEASURES ANALGESIC STUDY

Guoyong Jiang*, Cephalon, Inc.
Lilliam Kingsbury, Cephalon, Inc.

Randomized, breakthrough cancer pain studies involving two or more treatments often employ a repeated measures design in which each patient receives each treatment a fixed number of times in one of a prespecified subset of treatment sequences. To compare treatment means in such a design, we derive in this paper the uniformly most powerful invariant test. We illustrate the approach by applying it to real data from an analgesic study.

email: jjiang@cephalon.com

MIXED-EFFECTS PROBIT MODEL FOR LONGITUDINAL DATA WITH MULTIPLE DISCRETE OUTCOMES

Robert Gibbons, University of Illinois at Chicago
Hua Yun Chen*, University of Illinois at Chicago
Dullal Bauhmik, University of Illinois at Chicago

We propose a multivariate probit model for repeatedly measured multiple binary outcomes. We model the serial correlation at different occasions by random effects and use a factor analytic approach to model the correlation among multiple outcomes measured at the same occasion. The proposed model can be straightforwardly generalized to handle mixed discrete and continuous outcomes for repeatedly measured data by the latent variable argument of Shi \& Lee (2000). Gauss-Hermite quadrature approximations and/or Monte Carlo simulation are proposed to solve the computational problem in maximization of the marginal likelihood. The proposed methods are applied to the study of the childhood mental disorders.

email: hychen@uic.edu

MARGINALIZED REGRESSION MODELS FOR LONG SERIES OF LONGITUDINAL BINARY RESPONSE DATA

Jonathan S. Schildcrout*, Vanderbilt University
Patrick J. Heagerty, University of Washington

Marginalized regression models permit likelhood-based marginal regression inference for longitudinal categorical data. To ensure valid inference for regression parameters corresponding to both time-varying and time-invariant covariates proper specification of within-subject response dependence is usually required. While marginalized latent variable models (MLVM; Heagerty, 1999) capture exchangeable dependence and marginalized transition models (MTM; Heagerty, 2002) capture serially decaying dependence, it is common to observe both of these dependence model features when analyzing longitudinal series with many observation times. In this talk, we discuss an extension to the class of marginalized models that combines the MLVM and the MTM, thus permitting valid inference in the presence of serial plus exchangeable dependence. We discuss maximum likelihood estimation and the implications of dependence model misspecification. Specifically, we address the role that the covariate distribution can have on the validity of estimates in the presence of a misspecified dependence model. We also outline a Bayesian estimation strategy, and we conclude with an example.

email: jonathan.schildcrout@vanderbilt.edu

---

55. QUANTITATIVE-TRAIT LINKAGE ANALYSIS

A JOINT MODEL FOR NONPARAMETRIC FUNCTIONAL MAPPING OF GROWTH CURVES AND TIME-TO-EVENTS

Min Lin*, University of Florida
Rongling Wu, University of Florida

The characterization of the relationship between a longitudinal response process, such as growth curve, and a time-to-event has been a pressing challenge in biostatistical research. This has emerged as an important issue in genetic study when one attempts to detect the common genes or quantitative trait loci (QTL) that govern both growth trajectories and a developmental event such as the time to first flower in plants or the time for a tumor to reach a threshold size. In this article, we present a joint statistical model in which the event times and growth curves are taken to depend on a common set of genetic mechanisms. By fitting the Legendre polynomial of orthogonal properties for the time-dependent mean vector, our model presented here does not rely on any curve, which is different from our earlier parametric models. Our model allows for the detection of specific QTL that govern both growth and developmental processes through either pleiotropic effects or close linkage, or both. An example from a forest tree was used to demonstrate the usefulness of our model. The implications of this model for integrating vegetative growth and sexual reproduction to gain better insights into comprehensive biology are discussed.

email: mlin@mail.ifas.ufl.edu

VARIABLE SELECTION FOR LARGE P SMALL N REGRESSION MODEL WITH INCOMPLETE DATA: APPLICATION TO QTL MAPPING

Min Zhang*, Cornell University
Dabao Zhang, University of Rochester Medical Center
Martin T. Wells, Cornell University

Polygenic effects on complex traits are usually studied by collecting large number of molecular markers with relatively small sample sizes. The neighboring markers on the same chromosome are highly correlated and many marker data are missing due to failure in genotyping or selective genotyping. To map multiple Quantitative Trait Loci (QTL), we present a Bayesian approach to select variables for multiple linear regression with ``large p, small n' and incomplete data. The most important characteristic of the parameter space is that, though high-dimensional, it is sparse, which is incorporated into the prior with a probability mass concentrated at zero. To account for the different numbers of positive and negative coefficients, the prior is further specified as a mixing distribution with truncated normal distributions. In addition, our approach can naturally handle missing data. The inference was carried out by Markov chain Monte Carlo (MCMC) sampling scheme. The performance of the approach is evaluated by application to a publicly available dataset and by simulation study.

email: mz35@cornell.edu

ROBUST SEMIPARAMETRIC MULTIPOINT QUANTITATIVE-TRAIT LINKAGE ANALYSIS IN GENERAL PEDIGREES

Guoqing Diao*, University of North Carolina at Chapel Hill
Danyu Lin, University of North Carolina at Chapel Hill

The most commonly used quantitative trait loci (QTL) mapping methods in human genetics pertain to the variance component (VC) approaches. The conventional VC approaches extract the genetic attribution to the phenotypic variation from the identity-by-descent (IBD) allele-sharing information by fitting a linear mixed model, assuming that the quantitative trait follow a normal distribution, possibly after a known transformation. However, the transformation is usually unknown in practice and incorrect specification of the transformation may lead to biased parameter estimators and loss of power. To overcome these limitations, we extend the traditional VC approach to allow for a completely unspecified form of transformation. While retaining all the attractive features of traditional VC methods, the proposed approach is robust in that it does not require a parametric form of the transformation and that it is not sensitive to outliers. Extensive simulation studies demonstrate that the proposed method performs well in practical situations. Applications to a real study are provided.

email: gdiao@bios.unc.edu

STRUCTURED ANTEDEPENDENCE MODELS

Wei Zhao, University of Florida
Wei Hou, University of Florida
Ramon Littell, University of Florida
Rongling Wu, University of Florida

In this article, we present a statistical model for mapping quantitative trait loci (QTL) that determine growth trajectories of two correlated traits during ontogenetic development. This model is derived within the maximum likelihood context, incorporated by mathematical aspects of growth processes to model the mean vector and by structured antedependence (SAD) models to approximate time-dependent covariance matrices for longitudinal traits. The method incorporates logistic growth curve and SAD correlation structure to model growth process and correlation between different stages of growth. It allows for the test of the relative contributions of two mechanisms, pleiotropy and linkage, to genetic correlations. This model has been employed to map QTL affecting stem height and diameter trajectories in an interspecific hybrid progeny of Populus, leading to the successful discovery of three QTL on different linkage groups.

email: wzhao@stat.ufl.edu

USING GENERALIZED ESTIMATING EQUATIONS IN A GENOME SCAN OF CELL COUNTS IN THE DENTATE GYRUS OF RECOMBINANT INBRED MICE

Dirk F. Moore*, University of Medicine and Dentistry of New Jersey

The dentate gyrus is a structure in the hippocampus region of the brain, and recombinant inbred (RI) mice provide an ideal animal model for studying the genes that affect its characteristics. Methods for analyzing quantitative trait loci (QTL's) in inbred strains are well-developed, but data from RI mice present special opportunities and challenges. In particular, the availability of sets of multiple genetically identical mice with different combinations of genes from founders requires repeated measures methods such as those provided by generalized estimating equations (GEE's). In this talk I will describe a generalization of QTL methods to accommodate repeated measures of cell counts, and apply it to dentate gyrus data.

email: mooredf@umdnj.edu

A STATISTICAL FRAMEWORK FOR FUNCTIONAL MAPPING OF INTRACELLULAR CIRCADIAN RHYTHMS

Tian Liu*, University of Florida
Fei Long, University of Florida
Rongling Wu, University of Florida

Most living organisms display spontaneously sustained oscillations with a period close to 24 hours. Entrained to the daily temperature and light/dark cycles of the rotating earth, such circadian rhythms are regulated by a suite of genes that exhibit an interactive web. In this talk, we present a statistical framework for genomewide mapping specific genes (i.e., quantitative trait loci or QTL) that are responsible for variation in circadian rhythms at the intracellular organization level. Our framework is founded on statistical modelling of the mean- covariance structures of longitudinal traits within the context of a likelihood function constructed by finite mixture models. We capitalize on well-established mathematical models to fit the genetic regulations of various circadian processes at different time points and frame a series of stochastic processes to fit the structure of the covariance matrix. Using oscillations involved in mRNA and protein synthesis as an example, extensive simulation studies are performed to investigate the statistical robustness and behavior of our model for characterizing the rhythm QTL over a wide range of parameter values.

email: tianliu@ufl.edu

PRESIDENTIAL INVITED ADDRESS

SELECTION AND ESTIMATION FOR LARGE-SCALE SIMULTANEOUS INFERENCE

Bradley Efron*, Stanford University

Modern scientific technology is providing a new class of simultaneous inference problems for the applied statistician, where there are hundreds or thousands or even more hypothesis tests to consider at the same time. Microarrays epitomize this type of technology, but similar problems arise in proteomics, time of flight spectroscopy, flow cytometry, and functional Magnetic Resonance Imaging. I will consider two related questions: given a large number of simultaneous hypothesis testing situations, how can we Select the Non-Null cases; and how can we Estimate effect sizes for the Non-Nulls? The talk will use microarray data to illustrate a simple methodology for answering these questions.

email: brad@stat.stanford.edu

56. SCREENING FOR DISEASE: ISSUES IN STUDY DESIGN AND ANALYSIS

ESTIMATING THE ACCURACY OF CT COLONOGRAPHY

Valerie L. Durkalski*, Medical University of South Carolina
Yuko Palesch, Medical University of South Carolina
Peter B. Cotton, Medical University of South Carolina

Colorectal cancer is the 2nd most common cause of cancer- related deaths in the United States. It causes symptoms in only 10-15% of patients early in the disease, and therefore screening is necessary for early detection. Screening requires testing asymptomatic individuals for the presence of pre-malignant adenomatous polyps or colorectal cancer. Although procedures are available, less than 50% of patients over the age of 50 years undergo the standard screening method, conventional colonoscopy. A multicenter clinical study was designed to assess the accuracy of a new, less invasive, screening procedure (CT Colonography) for the detection of colon polyps. Nine centers collaborated in a study designed to compare CT colonography with conventional colonoscopy in the detection of colonic polyps and masses. This study is presented to illustrate study design and analytic issues encountered when evaluating screening tests. Challenges include defining the true presence/absence of disease or precursors of disease, accrual of patients with polyps (low prevalence), clustered data (more than one outcome per participant) and standardized clinical procedures across centers.

email: durkalsv@musc.edu

CHALLENGES TO STUDIES OF SCREENING TESTS WITH ILLUSTRATIONS FROM TOTAL BODY SCREENING WITH CT

Nancy A. Obuchowski*, Cleveland Clinic Foundation

The goal of screening is to detect disease before symptoms manifest and, sometimes, to prevent disease (e.g., removal of colon polyps). Designing studies to assess the efficacy of a screening program is complicated by the low prevalence of preclinical disease, the extended follow-up needed to observe patient outcome, preclinical conditions that will never develop into clinical disease (i.e., pseudodisease), and negative effects of the test itself (e.g., radiation, discomfort). Studies of the accuracy of screening tests are often hindered by the absence of reference tests and large reader variability. Even RCTs of screening programs are hindered by lead-time, length, and overdiagnosis biases, and cross-overs. Total body screening with Computed Tomography (CT) is a popular, highly publicized test with no scientific studies of its safety or effectiveness. We use total body screening with CT to illustrate these challenges to evaluating screening and offer some suggestions.

email: nobuchow@bio.ri.ccf.org

ROC CURVE EVALUATION OF SCREENING MODALITIES

Constantine Gatsonis*, Brown University
Mei Hsiu Chen, University of California, San Francisco

The evaluation of the diagnostic performance of tests in the screening context often leads to substantial requirements in terms of sample size and resources. To address these challenges, two-phase designs and Bayesian approaches have been proposed in the literature and have been developed primarily for the evaluation of binary screening tests. In this presentation we will discuss similar methods for ROC analysis of tests used in screening for disease. In particular, we will discuss both maximum likelihood and Bayesian formulations of the problem of estimating and comparing ROC curves using two-phase sampling designs. In many practical settings, these two-phase methods will be shown to provide more efficient alternatives to the usual single phase designs.

email: gatsonis@stat.brown.edu

57. COMBINING INFORMATION ACROSS SPATIAL SCALES

MODELING GLOBAL COVARIANCE STRUCTURES USING LOCAL INFORMATION

Petrutza C. Caragea*, Iowa State University

Environmental networks monitoring air pollutants often span over large geographical areas and include a very large number of locations. Modeling spatial correlation in a classical geostatistical sense could present several challenges, if one relies on the maximum likelihood estimation of the spatial parameters characterizing the 'global' covariance structure: restrictions on the computational capability when the number of spatial locations is very large as well as data collection patterns imposed by the general topography (less monitors located in mountainous regions, more sites located in highly populated areas). We propose a method of constructing a 'global' likelihood function that incorporates 'local' information through a conditional approach. This technique relies on the fact that the spatial process behaves as a stationary process at the local level, but not necessarily over the whole spatial domain. Another advantage of this method is that the computational effort associated with the estimation process is significantly diminished. Estimators obtained through the newly constructed 'global' model are compared to the 'local' estimators via a series of likelihood ratio tests. For data structures that allow the evaluation of the classical likelihood function, we compare the performance of the proposed model with the classical approach.

email: pcaragea@iastate.edu

MODELING NONHOMOGENEOUS POISSON PROCESSES USING A COMBINATION OF POINT AND COUNT DATA

Mark S. Kaiser*, Iowa State University
Han Wu, Iowa State University

Problems that involve the occurrence of certain events in a spatial domain may often be thought of as arising from an underlying nonhomogeneous Poission process. In some cases, data constitute the exact locations of all events. In other cases, data constitute aggregated counts of the number of events in various tracts contained in the spatial domain of interest. The situation considered here occurs when the data consist of a combination of these forms. Counts are available for each of a set of tracts that partition the spatial domain, and exact locations are available for a subset of the events of interest. For example, one may have the number of disease cases for each of a set of geographic subdivisions from a disease registry, and residential locations for a subset of cases treated at a particular medical center. We consider model formulation, estimation, and inference in such settings when there is also a spatially varying covariate available. The methods developed are illustrated using a data set of housing densities in the rural midwest.

email: mskaiser@iastate.edu

COMBINING DATA SOURCES TO EVALUATE SPATIAL AND TEMPORAL PATTERNS OF AVIAN POPULATION CHANGE

William A. Link*, USGS Patuxent Wildlife Research Center
John R. Sauer, USGS Patuxent Wildlife Research Center

We consider the problem of estimating range-wide population trends for bird populations. The North American Breeding Bird Survey (BBS) and the Christmas Bird Count (CBC) are typical of many bird surveys in having large geographic coverage which nonetheless is not coextensive with the range of certain species of management interest. Species' ranges contract and expand, and population distributions change through time, so that local rates of population change can be poor indicators of total population size. Combining data from various surveys is thus a necessity. Differences among surveys require care in producing composite analyses, but also provide opportunities for obtaining insights into species dynamics. In particular, differences in timing of surveys can sometimes be exploited to estimate temporal aspects of population change which cannot be estimated from individual surveys. We illustrate by examining seasonal effects on population change for Carolina wrens (Thryothorus ludovicianus) using CBC and BBS data.

email: william_link@usgs.gov

58. BAYESIAN PROCEDURES FOR ANALYZING MICROARRAY DATA

COMPARATIVE GENOMICS ANALYSIS OF GENE REGULATION

Jun Liu*, Harvard University
Cristian Castillo-Davis, Harvard University
Lei Shen, Harvard University

With the availability of an increasing number of species' genomes and the explosive amount of gene expression data, it becomes apparent that one can gain critical information by comparing genomes of different species, and a great deal of insight by combining the comparative genomics information with gene expression profiling information. I will describe some of our recent work on finding regulatory motifs or modules and studying regulatory sequence evolution based on comparing human, mouse, and other eukaryotic genomes.

email: jliu@stat.harvard.edu

TOWARDS A COMPLETE PICTURE OF GENE REGULATION: USING BAYESIAN APPROACHES TO INTEGRATE GENOMIC SEQUENCE AND EXPRESSION DATA

Mayetri Gupta*, University of North Carolina at Chapel Hill
Joseph G. Ibrahim, University of North Carolina at Chapel Hill

Over the last decade, the profuse generation of genomic data through whole-genome sequencing and advances in gene expression microarray technology has led to a spurt in statistical research in an effort to make biologically meaningful predictions. The analysis of sequence and expression data have often been considered two separate problems, though biologically, they are intrinsically related. Regulatory motif discovery tends to focus on promoter sequences of genes that are believed to be co-regulated. The behavior of multiple genes in an organism is studied through analyzing mRNA expression from whole-genome microarrays, ignoring the genomic sequence. This approach may result in missing potentially important regulatory motif interactions or the occasional construction of clusters based on artifacts that have no biological significance. In this talk we propose an encompassing model that captures biologically plausible relationships between genomic sequence features and resultant gene expression. Using this model, we find clusters of genes that show evidence of combinatoric regulation, simultaneously inferring motifs that may have combinatorial effects within clusters, or differential effects across clusters. The applicability and preliminary successes of our methods are demonstrated using real biological examples, and extensions of our method towards integrating more complex data sources such as chIP experiments are indicated.

email: gupta@bios.unc.edu

BAYESIAN VARIABLE SELECTION METHODS FOR THE ANALYSIS OF DNA MICROARRAY DATA

Mahlet G. Tadesse*, University of Pennsylvania
Naijun Sha, University of Texas at El Paso
Marina Vannucci, Texas A&M University

The identification of distinct subtypes of a disease and the detection of discriminating genes are common goals in the analysis of DNA microarray data. Molecular classes defined on a small number of genes can lead to a better understanding of the underlying biological processes. In addition, the selected genes can serve as biomarkers for improved diagnosis and targets for therapeutic intervention. In this talk, I will present Bayesian variable selection methods in the context of classification and clustering. Although somewhat related, these two problems are quite different. In classification, the group structure is specified in a training dataset and guides the variable selection. In clustering, however, the outcome is not observed and the analysis is fully data-based. This calls for different methods for modeling the data and implementing the Bayesian variable selection searches.

email: mtadesse@cceb.upenn.edu

59. COST-EFFECTIVENESS ANALYSIS: METHODOLOGIES FOR COMPARING COMPETING HEALTH CARE INTERVENTIONS

DESIGNING NATIONAL HEALTH CARE EXPENDITURE SURVEYS TO INFORM HEALTH POLICY AND PRACTICE

Steven B. Cohen*, AHRQ

Health care expenditures represent nearly one-seventh of the United States gross domestic product, exhibit a rate of growth that exceeds other sectors of the economy, and constitute one of the largest components of the Federal and states' budgets. In order to allow for comprehensive studies of the current health care system, information is needed on the population's access to health care, their utilization of and expenditures for health care services and their health insurance coverage. To effectively address these issues, researchers and policymakers need accurate nationally representative data to better permit an understanding of how individual characteristics, behavioral factors, financial and institutional arrangements affect health care utilization and expenditures in a rapidly changing health care market. The demand for accurate and reliable information on the population's health care expenditures, insurance coverage and sources of payment is met by the Medical Expenditure Panel Survey (MEPS) sponsored by the Agency for Healthcare Research and Quality (AHRQ). In this presentation, the design features of the MEPS are presented with an emphasis of those components that have been implemented to inform health policy and practice with respect to health care costs.

email: scohen@ahrq.gov

QOL ADJUSTMENTS FOR NONDEGRADATION PROCESSES IN LIFE TIME ANALYSIS

Pranab K. Sen*, University of North Carolina at Chapel Hill

In a stochastic environment, a degradation process, in spite of showing a monotone trend, may contain some stochastic variations which may camouflage the statistical picture to a certain extent. There are, however, some other processes arising in some chronic diseases which may not exhibit a degradation phenomenon. For some of these nondegradation stochastic processes, associated aging perspectives are appraised: the scenario beyond semiparametrics is examined in this context with due emphasis on health related quality of life assessment. Counting processes are thoroughly assessed in this respect.

email: pksen@bios.unc.edu

A DYNAMIC MODEL TO ASSESS COVARIATE EFFECTS ON COST AND HEALTH OUTCOMES

Joseph C. Gardiner*, Michigan State University
Zhehui Luo, Michigan State University
Corina M. Sirbu, Michigan State University
Cathy J. Bradley, Michigan State University
Charles W. Given, Michigan State University

Patients undergo changes in health status as their event history unfolds over time. Costs are incurred through resource use while patients sojourn in health states. Because complete observation of these transitions and sojourns is not always feasible, many patient histories will be incomplete and total costs will be unknown. To address this problem, we develop a dynamic model for assessing covariate effects on transitions between health states using a Markov model to estimate the transition probabilities. Next, we use a random-effects model for sojourn costs with transition times as random effects. The two models are combined to estimate net present values of expenditures over a finite time interval as a function of patient characteristics. The method is applied to a data set of 624 incident cancer cases. Changes in physical functioning over 30 months are described by a 3-state Markov process (low or normal physical function, or dead). There were 588 completed transitions. Charges for inpatient, outpatient and physician services were derived from administrative records for a 24 month period. We estimate net present values for charges incurred over 2 years by cancer stage. The joint regression model provides a flexible approach to assessing the influence of patient characteristics on both cost and health outcomes while accommodating heteroscedasticity, skewness and censoring in the data.

email: jgardiner@epi.msu.edu

SEMIPARAMETRIC SPATIAL MODELING OF BINARY OUTCOMES

Tatiyana V. Apanasovich*, Cornell University
David Ruppert, Cornell University
Raymond J. Carroll, Texas A&M University

Important features of the motivating example include a large number of repeated measures per subject and a small number of subjects. Many papers on longitudinal data address data sets of the opposite type where there are many subjects but relatively few repeated measures per subject. Our aim is to produce a theoretical framework for longitudinal and spatial data which is general enough to apply to our example as well as many others with a large number of outcomes per subject. Such examples typically exhibit several features: (1) nonlinearity may be seen so that a nonparametric model of the mean function, that is, the conditional expectation of the response given the covariates, is likely to be needed, and (2) correlation between repeated measures can be estimated sufficiently accurately that nonstationarity may be evident. Our model of the mean function is new and includes single-index models, additive models, and partially linear models as special cases. Though motivated by a spatial example, it is not restricted to longitudinal and spatial data.

email: tanya@orie.cornell.edu

TRANSFER FUNCTIONS IN HIERARCHICAL FUNCTIONAL DATA MODELS, WITH APPLICATIONS TO PREDICTING CELL PROLIFERATION AND APOPTOSIS FROM P27 EXPRESSION IN COLON CARCINOGENESIS EXPERIMENTS

Raymond J. Carroll*, Texas A&M University
Veera Baladandayuthapani, Texas A&M University
Kimberly Drews, Texas A&M University

In colon carcinogenesis experiments, p27 expression is thought to be predictive of cell proliferation and apoptosis (programmed cell death). In order for a cell to either proliferate or undergo apoptosis, it must first receive the signals to do so and respond accordingly. Therefore, as a cell moves up the colonic crypt, there is a time component built into the position the cell holds in the crypt. This suggests the conjecture that it is p27 expression in the cell as well as the cells below it (younger cells) that affect proliferation and apoptosis. We investigate this conjecture using data from a recently completed experiment at Texas A&M. The data arise naturally as hierarchical functional data, with multiple animals having multiple colonic crypts and multiple cells within the crypts. In order to answer the question posed by our biologists, we formulate a novel combination of hierarchical regression splines together with transfer function models. The key point here is that inference is meant to be at the crypt level, rather than at the rat level, exactly the opposite of the aim of most hierarchical analyses. Two general approaches are investigated. In the first, we marginalize to the rat level to eliminate the nuisance parameters that would ordinarily be of interest. The second approach is Bayesian: instead of brute-force MCMC, we marginalize to improve the mixing of the chain.

email: carroll@stat.tamu.edu

SPATIALLY ADAPTIVE BAYESIAN P-SPLINES WITH HETEROSCEDASTIC ERRORS

Ciprian M. Crainiceanu*, Johns Hopkins University

Nonparametric smoothing using fixed-knot penalized spline (P-splines) is a very powerful regression tool. Its success is based, in part, on the use of low-order spline bases which make computations tractable while the accuracy is as good as with smoothing splines. The standard approach is to choose a relatively small number of basis functions (knots) and use a global penalty to avoid oversmoothing. We extend this methodology by allowing the error variance and the penalty parameter to vary smoothly over time. Modeling the error variance nonparametrically improves estimation efficiency and describes how variability changes with the predictor. An example of intrinsic interest is when a treatment affects not only the mean but also the variance of the response. Modeling the penalty parameter improves estimation efficiency when the mean function is slowly varying in some regions of the space and rapidly varying in others. A fully Bayesian approach is used to provide the joint posterior distribution of the model parameters. In particular, we obtain the posterior distributions of the error standard deviation and penalty functions. An important feature of our methodology is that is can be quickly implemented using the Bayesian software WinBUGS. We discuss the important problem of prior choice in the context of P- spline smoothing and extensions to multivariate smoothing.

email: ccrainic@jhsph.edu

A SEMIPARAMETRIC MODEL FOR A NONSTATIONARY TIME SERIES OF COUNTS WITH TIME-DEPENDENT COVARIATES

Andres Houseman, Harvard University
Brent A. Coull*, Harvard University
James P. Shine, Harvard University

We propose a negative binomial model for a time series of counts when interest focuses on the effect of time- dependent covariates. The model accounts for nonstationarity and autocorrelation using a nonparametric smooth function of time. We discuss identifiability issues related to estimating the effects of time-dependent covariates in the presence of a nonparametric function of time, and propose a penalized spline approach to estimation based on a Fourier basis that ensures identifiability of all effects of interest. We apply the model to eight years of water pathogen data from a Boston Harbor monitoring study.

email: bcoull@hsph.harvard.edu

SEGMENT RESPONSE SURFACE FOR SYNERGY ANALYSIS

Xiaoli S. Hou*, Merck & Co., Inc.
Keith A. Soper, Merck & Co., Inc.

A new "segment response surface" approach is proposed for measuring the joint effect of two or more drugs in vitro. Commonly used procedures parameterize how drugs interact, usually employ a pre-specified model or assuming the same degree of synergy or antagonism. Results can be misleading when the true model is different from the pre-specified. Nonparametric methods lack power and are often difficult to summarize. Our method divides the observed dose-space into smooth regions where synergy, additivity, and antagonism are observed. In addition, we subdivide "antagonism" into clinically important regions. "Antagonism" at some dose combination of drugs A and B can consist of a response more beneficial than that for either drug alone, though less than that predicted by additivity; the response can be equal to that for drug A given alone; or the response can be worse than that for drug A alone. Our "segment response surface" method yields intuitive visual summaries, simplifies modelling, and can be extended directly to combinations of more than two drugs. Associated statistical tests are more powerful than competing standard methods in realistic situations. Finally, we discuss how to detect and partially adjust for systematic measurement error such as row or column effects that sometimes occur on plate(s) used to run the assay.

email: xiaoli_hou@merck.com

TESTING IMMUNOLOGICAL CORRELATES OF PROTECTION

Andrew J. Dunning*, Wyeth Vaccines Research

Immunological assays measure characteristics of the immune system, such as antibody levels, specific to certain diseases. High assay values are often associated with protection from disease. Previous research has proposed a scaled logit model to quantify the relationship between assay values and protection from disease; the model was set in the context of an undifferentiated population. A question of interest is whether the relationship between assay values and protection from disease differs in sub- populations. Of particular interest in the context of vaccine research is whether the relationship between assay values and protection from disease is the same for vaccinees and non-vaccinees. Also of interest is whether the relationship between assay values and protection can be shown to be the same in different clinical trials. Inferential methods will be presented for evaluating these questions and illustrated with examples.

email: adunning@alumni.washington.edu

ASSESSING IN VITRO BIOEQUIVALENCE FOR PROFILE DATA: A NEW MODELING APPROACH

Bin Cheng*, Columbia University

To establish the bioavailability (BA) and/or the bioequivalence (BE) of some locally acting drugs such as nasal aerosols and nasal sprays, the 2003 FDA guidance suggests that a profile analysis of the particle size distribution (PSD) by cascade impactor be included in the in vitro studies. Due to the multivariate non-normal nature of the data, an appropriate method to assess the BE between the test and reference drugs is not available although several approaches have been suggested. To address the disadvantages of the existing methods, we propose a new approach to assessing the in vitro BE by modeling the transformed PSD profiles.

email: bc2159@columbia.edu

REDUCING ANIMAL USE IN CHEMICAL TOXICITY TESTING

Robert Lee*, Constella Health Sciences
Eric Harvey, Constella Health Sciences
Patrick Crockett, Constella Health Sciences
Shyamal Peddada, National Institute of Environmental Health Sciences

Before chemicals are approved for commercial use, animal toxicity studies are performed to classify the chemicals into a toxicity category. Our focus is to classify the chemicals using as few animals as possible. The current standard procedure used by the EPA is a sequential dosing procedure called the Up and Down Procedure (UDP) (EPA Health Effects Test Guidelines, 2002). Toxicity studies are also performed in Phase I clinical trials, where it is critical to use the fewest number of subjects possible. Methods developed for this situation include a Bayesian method called the Continual Reassessment Method (CRM) by O'Quigley et al. (Biometrics, 1990) and a more recent nonparametric method using order-restricted inference developed by Conaway et al. (Biometrics, 2004). Phase I techniques may be used in animal studies to effectively classify chemicals into toxicity categories and limit the number of animals used. We explain how both CRM and the nonparametric method can be used in animal testing studies. Simulated data are used to compare these two methods to the UDP. Comparisons are based on the number of animals used and the accuracy of classification into toxicity categories.

email: eharvey@constellagroup.com

STATISTICAL ANALYSIS OF CDFSE DATA

Ollivier Hyrien*, University of Rochester Medical Center

CDFSE data are frequently used in biology to gain quantitative insight into cell population kinetics. Immunologists use extensively this technique to better understand the dynamics of immune responses. These data allow the experimentalist to track the number of cell divisions undergone from the beginning of the experiment. To date, only a limited number of methods have been proposed for the statistical analysis of CDFSE data. During this talk, we will describe a new method based on a modeling approach. In the proposed method, a branching process represents cell proliferation, and random variables attached to all cells of the population describe the evolution of the level of CDFSE over time. We present a procedure for fitting this model, and apply the proposed method to a set of experimental data.

email: ollivier_hyrien@urmc.rochester.edu

A RESPONSE SURFACE MODEL FOR DRUG COMBINATIONS

Maiying Kong*, University of Texas M. D. Anderson Cancer Center
J. Jack Lee, University of Texas M. D. Anderson Cancer Center

When drugs having like effect are given in combinations, investigators often want to assess whether the joint effect is additive, synergistic, or antagonistic. Several response surface models (Machado and Robinson 1994; Greco et al. 1990; Carter et al. 1988; Plummer and Short 1990) have been proposed in the literature by using a single parameter to capture synergism or antagonism. Limitation of these models exists when combinations at certain doses are synergistic while at other doses of the exact same drugs are antagonistic. We propose a response surface model, where a function of doses in a combination instead of a single parameter will be used to identify and quantify departures from additivity for different combinations. The proposed model can be considered as a generalized form of Plummer and Short's model (1990), and could capture all forms of synergism, antagonism, and additivity at different combinations of the two drugs. Examples will be given to illustrate the advantages of using the proposed method.

email: mykong@mdanderson.org

COMPARING SYNERGY

Gregory D. Ayers*, University of Texas M. D. Anderson Cancer Center
J. Jack Lee, University of Texas M. D. Anderson Cancer Center

The principle of Loewe additivity has emerged as the standard for evaluating synergy, antagonism, or additivity among two or more agents (Greco 1995). Several statistical models, based either on mechanistic or empirical considerations, are in common use to test for synergistic/antagonistic drug interactions. We compare and contrast several popular models including, 1) the median effect model (Chou and Talalay 1984), 2) the Greco model (Greco et al 1990), and 3) general estimation of the joint action ratio from a class of models (Machado and Robinson 1994). We analyze and summarize data from 10 cell lines of human squamous cell carcinoma of the head and neck treated with two pairs of biologic agents. The combination index, alpha, and eta (joint action ratio) from models 1, 2, and 3, respectively, are single parameter estimates that indicate synergy or antagonism. 95% confidence intervals containing 1 for model 1 and 3, or containing 0 for model 2, indicate lack of evidence to infer non-additivity. We observed appreciable lack of consensus among these models for the same cell lines. Causes for these differences will be explored; geometric interpretation, experimental design, and other issues will be discussed.

email: danayers@mdanderson.org

62. SEQUENTIAL METHODS

A SEQUENTIAL TEST FOR TREATMENT EFFECTS UNDER STAGGERED ENTRY IN A MULTICENTER CLINICAL TRIAL

Dong-Yun Kim*, Illinois State University
Michael B. Woodroofe, University of Michigan

Consider a multicenter clinical trial where patients enter for treatment at random times; upon arrival, each patient is randomly assigned to one of the two treatments. We develop a sequential probability ratio test that compares the effects of treatments on patients' survival distribution while allowing for 'site' effects. We show that the test statistic, viewed as a stochastic process indexed by the sample size, can be approximated by a random walk perturbed by a stationary sequence and a slowly changing sequence. We prove a non-linear renewal theorem for the perturbed process and derive the asymptotic joint distribution of the perturbations and the excess at the first passage time. Using Monte Carlo simulation, we determine the critical values of the test from the excess distribution.

email: dongyunhkim@yahoo.com

SUPPLEMENTARY ANALYSIS OF PROBABILITIES AT THE TERMINATION OF A GROUP SEQUENTIAL PHASE II TRIAL

Aiyi Liu*, National Institute of Child Health and Human Development
Chengqing Wu, National Institute of Child Health and Human Development
Kai F. Yu, National Institute of Child Health and Human Development
Edmund A. Gehan, Georgetown University Medical Center

We consider estimation of various probabilities after termination of a group sequential phase II trial. A motivating example is that the stopping rule of a phase II oncologic trial is determined solely based on response to a drug treatment, and at the end of the trial estimating the rate of toxicity and response is desirable. The conventional maximum likelihood estimator (sample proportion) of a probability is shown to be biased, and two alternative estimators are proposed to correct for bias, a bias-reduced estimator obtained by using Whitehead's bias-adjusted approach, and an unbiased estimator from the Rao-Blackwell method of conditioning. All three estimation procedures are shown to have certain invariance property in bias. Moreover, estimators of a probability and their bias and precision can be evaluated through the observed response rate and the stage at which the trial stops, thus avoiding extensive computation.

email: liua@mail.nih.gov

EXACT GROUP SEQUENTIAL DESIGNS FOR DETECTING HYPOTHESES INVOLVING TWO CORRELATED BINARY RESPONSES

Jihnhee Yu*, Roswell Park Cancer Institute
James L. Kepner, Roswell Park Cancer Institute
Brian N. Bundy, Roswell Park Cancer Institute

A 1-stage exact design to test joint hypotheses involving success rates is discussed using a new way to model bivariate binomial variables. The method is extended to obtain two-stage exact group sequential designs. Two types of two-stage designs are presented, one is similar to Bryant and Day's two-stage design, and the other is analogous to Kepner and Chang's Type I design. The results demonstrate how the required sample sizes to assure a targeting power and significance level vary as a function of the correlation.

email: yuj@roswellpark.org

COMBINATIONS OF TWO-STAGE DESIGNS FOR TESTING MULTIPLE TREATMENTS IN PHASE II CANCER TRIALS

Tatsuki Koyama*, Vanderbilt University School of Medicine
Jordan D. Berlin, Vanderbilt University School of Medicine
Yu Shyr, Vanderbilt University School of Medicine

In Phase II cancer trials, there is sometimes more than one candidate treatment, but a multi-arm trial may be logistically difficult to conduct. For such a situation, we propose an extension of a general single arm two-stage design to allow switching to a different candidate treatment at the end of each stage. The treatment of primary interest is tested first in a two-stage design, and the second two-stage design is conducted sequentially if the primary candidate treatment is found inactive in either stage I or stage II. The framework is extended to include the third two-stage design when there are three candidate treatments. Alternatively, the first two-stage design may have three decisions to incorporate an additional treatment. The choice of design depends on the assumptions we make about the response rates of the candidate treatments. In this presentation, we illustrate a design framework and examine the characteristics of various types of designs in terms of expected and maximum sample sizes and type I and II error probabilities.

email: tatsuki.koyama@vanderbilt.edu

BAYESIAN OPTIMAL DESIGN FOR PHASE II TRIALS

Meichun Ding*, Rice University
Gary L. Rosner, University of Texas M. D. Anderson Cancer Center
Peter Mueller, University of Texas M. D. Anderson Cancer Center

Most phase II screening designs available in the literature consider one treatment at a time. Each study is considered in isolation, even when prior data exist on the treatment or related therapies. We propose a more systematic and rational decision-making approach to the phase II screening process based on decision-theoretic Bayesian optimal design. The sequential decision problem is to choose one action out of three, based on what we have learned about the agent: stop to abandon the treatment, stop to switch to a phase III trial, or continue sampling. The design criterion is to maximize the utility for the new treatment, incorporating patient costs and potential payoff. The underlying probability model is a hierarchical probit regression model that also allows for covariates. The Bayesian hierarchical model allows combining information across several related studies in a formal way. In particular, prior knowledge of the success probabilities of related new treatments is sequentially updated as more patients enter the study and data accumulate. The proposed methodology is applicable in any setting in which one wishes to screen several contending innovations that appear over time.

email: meichund@stat.rice.edu

CONDITIONAL MAXIMUM LIKELIHOOD ESTIMATION FOLLOWING A GROUP SEQUENTIAL TEST

Aiyi Liu, DHHS/NIH/NICHD/DESPR
James Troendle, DHHS/NIH/NICHD/DESPR
Kai F. Yu, DHHS/NIH/NICHD/DESPR
Weishi Yuan*, DHHS/FDA

We consider estimation after a group sequential test. An estimator that is unbiased or has small bias may have substantial conditional bias. We derive the conditional maximum likelihood estimators of the parameters, and investigate their properties within a conditional inference framework. The method applies to both the usual and adaptive group sequential test designs.

email: yuanv@cber.fda.gov

ESTIMATING RATES OF RESPONSE AND TOXICITY FOLLOWING A BIVARIATE GROUP SEQUENTIAL PHASE II CLINICAL TRIAL

Chengqing Wu*, BMSB,DESPR, NICHD, NIH
Aiyi Liu, BMSB,DESPR, NICHD, NIH
Kai F. Yu, BMSB,DESPR, NICHD, NIH
Ming Tan, Greenebaum Cancer Center, University of Maryland

We consider estimation of probabilities after termination of a group sequential phase II trial with early stopping based on both response and toxicity observations. For each probability of interest, such as response or toxicity rate, the sample proportion continues to be the maximum likelihood estimator but often possesses substantial bias. We construct three alternative estimators, a bias-adjusted estimator using Whitehead's bias-adjustment approach, an unbiased estimator obtained from the Rao-Blackwell method, and a simple bias-reduced estimator by direct bias-adjustment. Numerical results show that three estimators reduce the bias substantially while maintaining satisfactory mean squared error.

email: wuch@mail.nih.gov

SEMIPARAMETRIC JOINT MODELING OF LONGITUDINAL MEASUREMENTS AND TIME-TO-EVENT DATA

Wen Ye*, University of Michigan
Xihong Lin, University of Michigan
Jeremy M. G. Taylor, University of Michigan

Longitudinal studies in medical research often generate both censored time-to-event data and repeated measurements on biomarkers. Recently, joint models using both types of data have been developed. Commonly, the longitudinal covariate is modeled by a linear mixed model. However, in some cases, the biomarker's time trajectory is not linear, such as the prostate specific antigen PSA profile after radio-therapy in prostate cancer study. We propose a two-stage regression calibration approach which models the longitudinal biomarker using a semiparametric mixed model, where covariate effects are modeled parametrically and the individual time trajectories are modeled nonparametrically using a population smoothing spline and subject-specific random stochastic processes. Estimates of the biomarker level and change rate at each time-event are then used as time dependent covariates in the second stage survival model. The performance of the approach is illustrated by application to a prostate cancer study.

email: wye@umich.edu

ESTIMATION IN TWO-STAGE RANDOMIZATION DESIGNS

Xiang Guo*, North Carolina State University
Anastasios A. Tsiatis, North Carolina State University

In many clinical trials related to diseases such as cancers and HIV, patients are treated by different combinations of therapies. This leads to two-stage designs, where patients are initially randomized to a primary therapy and then depending on the disease remission and patients' consent, a maintenance therapy will be randomly assigned. In such designs, the effects of different treatment policies, i.e., combinations of primary and maintenance therapy are of great interest. In this paper, we use the ideas of counterfactual random variables to define the survival time for various treatment combinations of induction and maintenance therapies and we also use concepts of counting process and risk sets to find weighted estimating equations whose solution gives an estimate for the cumulative hazard function which, in turn, is used to derive the estimator for the survival distribution with right- censored data. Our estimator is a natural extension of the Aalen-Nelson estimator for the cumulative hazard function. It is more intuitive and easier to compute and, as we will demonstrate, is more efficient than the available estimators

email: xguo4@stat.ncsu.edu

NONPARAMETRIC REGRESSION USING KERNEL ESTIMATING EQUATIONS FOR CORRELATED FAILURE TIME DATA

Zhangsheng Yu*, University of Michigan
Xihong Lin, University of Michigan

We study nonparametric regression for the correlated failure time data under the marginal proportional hazard model. Kernel regression estimating equations are used to estimate nonparametric covariate effects. Independent and weighted working kernel estimating equations (EE) derived from local partial likelihood are studied. The derivative of the covariate function is first estimated and the covariate function estimator is obtained by integrating the derivative estimator. The Trapezoidal rule is used for integration approximation. We show that the nonparametric estimator of the covariate function's derivative is consistent for any arbitrary working correlation matrix and the asymptotic variance is minimized by assuming working independence. We evaluate the performance of the proposed kernel estimator using simulation studies, and apply the proposed method to western Kenya parasitaemia.

email: zyuz@umich.edu

METHODS FOR ANALYZING SURVIVAL DATA FROM ALTERNATING STUDIES

Beth Ann Griffin*, Harvard University
Stephen Lagakos, Harvard University

We develop statistical methods for designing and analyzing survival studies in which treatments are deliberately interrupted and reinitiated, but where interest lies in making inferences about continuous treatment use. We refer to such designs as alternating designs since subjects alternate between periods in which they are taking the treatment of interest and periods when they are not. We examine a nonparametric estimator of the cumulative hazard function for continuous treatment and show it to be uniformly consistent and asymptotically normal under certain conditions relating to the effects of interrupting the treatment. We then introduce nonparametric tests for comparing the distributions corresponding to two such continuously-given treatments, derive their asymptotic properties under general alternatives to the null, and give conditions for their asymptotic validity. The properties of the alternating treatment design and the classical parallel group design are compared. We illustrate the proposed methods using the results from a recent study by Gallin et al. (2003) in which subjects alternate between taking an active drug and placebo on an annual basis.

email: bgriffin@hsph.harvard.edu

A SEQUENTIAL STRATIFICATION METHOD TO ESTIMATE THE EFFECT OF A TIME-DEPENDENT TREATMENT IN THE ANALYSIS OF RECURRENT EVENT DATA

Douglas E. Schaubel*, University of Michigan
Robert A. Wolfe, University of Michigan

Recurrent event data often arise in biomedical studies. We propose a semiparametric method for estimating the effect on the marginal event rate of a time-dependent treatment. When applied to observational studies, the proposed method is intended to yield parameter estimates which have more of a causal interpretation than those of existing methods. Asymptotic properties of the regression parameter estimator are derived and evaluated in finite samples through simulation. Data from the retrospective cohort study which motivated the proposed methods, from a national organ failure registry, will be analyzed using the proposed and previously existing methods.

email: deschau@umich.edu

SEMIPARAMETRIC TRANSFORMATION MODELS FOR THE CASE-COHORT STUDY

Wenbin Lu*, North Carolina State University
Anastasios Tsiatis, North Carolina State University

The case-cohort design was introduced by Prentice (1986) for large epidemiological and other event history studies. Under the case-cohort design, a random sample from the entire cohort is selected, named the subcohort. The covariate information is only collected for the subjects in the subcohort and any cases, subjects who experience the event of interest, outside the subcohort. Case-cohort designs are used for saving costs in many large medical studies, especially when events are usually rare and covariates are expensive. In this paper, a general class of semiparametric transformation models is studied for analyzing survival data from the case-cohort design. Weighted estimating equations are proposed for simultaneous estimation of the regression parameters and the transformation function. It is shown that the resulting regression estimators are asymptotically normal, with variance-covariance matrix that has a closed form and can be consistently estimated by the usual plug-in method. Simulation studies show that the proposed approach is appropriate for practical use.

email: lu@stat.ncsu.ed

MATRIX SKEWED DISTRIBUTIONS

Solomon W. Harrar*, South Dakota State University
Arjun K. Gupta, Bowling Green State University

Suppose p measurements are taken from n experimental units. Under usual assumption of iid multivariate normalilty several statistical inference procedures have been devised and used. Although such modeling of data appears to be realistic, the effect of the violations of the independence and normality assumptions need to be explored to reduce the risk resulting from wrong decisions. The main purpose of this paper is to develop a general model that is capable of modeling multivariate data by allowing skewness and kurtosis, and dropping the assumption of independence between the experimental units. The approach adopted involves adding a skewing factor to matrix variate elliptical models. Thus, the model provides a reasonable class of distributions for studying the robustness of many multivariate data analysis techniques. In particular, in this paper we derive the properties of matrix skew-elliptical distribution by placing emphasis on the matrix variate skew-normal and matrix variate skew-t distributions. Among others we study the properties of the linear and quadratic forms of random matrices with this distribution.

email: solomon.harrar@sdstate.edu

LONGITUDINAL DATA ANALYSIS IN F1-LD-F1 FACTORIAL DESIGN WITH APPLICATION TO LIVER CANCER STUDY

Ke Yan*, Texas Tech University

The objective of the research is to test the efficiency of green tea polyphonols (GTP) in reducing the concentration of aflatoxin B1 (AFB1) in blood samples. We apply the rank based method in the factorial design (F1_LD_F1) for longitudinal data. The hypotheses and test statistics are introduced for average time effect, simple time effect, group effect as well as interaction between time and group. Results of this research are likely to have an impact on improving the chemoprevention in liver cancer study.

email: ke.yan@ttu.edu

METHODS OF LONGITUDINAL DATA FOR F2-LD-F1 MODEL

Lan Zhang*, Texas Tech University

For multifactorial experiments with longitudinal data, the rank based method (nonparametric method) is considered where the hypotheses and test statistics are introduced for average group effect, average time effect, simple time effect, and the interaction between time and group. The goal of this paper is to provide a description of a nonparametric model, F2_LD_F1 model, to illustrate the fundamental ideas, and to demonstrate its application in green tea polyphnols (GTP) studies. Three strategies, missing values strategy, last observation carried forward strategy, and complete cases strategy are also used to deal with the missing data from this study.

email: lan.zhang@ttu.edu

CONTROLLING VARIABLE SELECTION BY THE ADDITION OF PSEUDO-VARIABLES

Yujun Wu*, University of Medicine and Dentistry of New Jersey
Dennis D. Boos, North Carolina State University
Leonard A. Stefanski, North Carolina State University

We propose a new and general approach to variable selection, designed to control the false selection rate (FSR), i.e., the proportion of unimportant variables included in the final model. The method works by adding a known number of pseudo-variables to the real data set, running a variable selection procedure, and monitoring the proportion of pseudo-variables falsely selected in the model. Information obtained from bootstrap-like replications of this process can be used to estimate the number of falsely-selected real variables and also to tune the selection procedure to control the false selection rate. We focus on forward selection because it is applicable in the case where there are more variables than observations. Due to the difficulty of obtaining analytical results, we study our approach by Monte Carlo, comparing it to a number of common variable-selection procedures. The new method is illustrated on a real example.

email: wuy5@umdnj.edu

LEARNING CURVE ANALYSIS OF LOGISTIC REGRESSION AND TREE-STRUCTURED ALGORITHMS

Qinghua Song*, University of Wisconsin–Madison
Wei-Yin Loh, University of Wisconsin–Madison
Kin Yee Chan, National University of Singapore, Republic of Singapore
Yu-Shan Shih, National Chung Cheng University, Taiwan, Republic of China

Logistic regression is a standard statistical approach to modeling binary data. Results from some previous comparison of linear logistic regression with C.5 show that the classification accuracy of linear logistic regression is better for small to moderate-sized datasets but that it does not outperform C4.5 when the sample size is large. This is due to the number of variables in logistic linear regression being held fixed whereas the number of nodes in a C4.5 model can grow with the sample size. In our investigation, we use learning curves to study the effectiveness of tree-structured models (C4.5, LOTUS, and QUEST) and logistic regression methods (linear, stepwise quadratic, and stepwise interaction) for classification and probability rankings. We find some patterns indicating that certain logistic regression methods can be extremely competitive even for large data sets.

email: songq@stat.wisc.edu

TRANSFORMATION SUPPORTING EDF GOODNESS-OF-FIT TEST OF NORMAL DISTRIBUTION RELATED TO TWO SAMPLES BASED ON REGRESSION

Dhanuja Kasturiratna*, Bowling Green State University
Truc T. Nguyen, Bowling Green State University
Arjun K. Gupta, Bowling Green State University

The one-way classification with two treatments, we assume that the data are observed according to the additive model, where the error random variables are independently and identically distributed normally with mean zero and unique variance. Then to test whether a set of observed data, coming from the above regression model, we need to construct a test to test the hypothesis that the observations of the two treatment groups are normally distributed with same variance. Here the means of two treatment groups and the common variance are unknown. The motivation of characterization given in this paper is to find a transformation in the procedure to construct an exact EDF goodness-of-fit test for testing the above hypothesis.We obtain characterization of normal distribution in two samples based on second conditional moments. This characterization has been changed to characterization based on the UMVU estimators of the density functions, then to characterization using Student's t distribution.Using the characterization results obtained, the above composite hypothesis can be changed to a simple hypothesis that can be tested using any EDF statistic, such as Kolmogorov-Smirnov statistic,Cramer-von Mises statistic, Anderson-Darling statistic etc. Finally these results are generalized for k samples.

email: dhanuja@bgnet.bgsu.edu

THE ROLE OF PERCENTILES IN DETERMINING A REGRESSION EQUATION

Yvonne M. Zubovic*, Indiana University Purdue University Fort Wayne
Chand K. Chauhan, Indiana University Purdue University Fort Wayne

Fitting a least squares line to data exhibiting a linear relationship between X and Y is a common practice. The effect of outlying observations on the least squares regression analysis has been well documented. Many alternatives have been suggested to reduce the influence of outliers and yield a robust regression line. One such method, the median-median line approach, has been discussed as a method of minimizing the effect of outliers. In this approach the data are partitioned into three segments based on the magnitude of the X values. The medians of both the X and Y values are computed for these segments. These medians are used to estimate the regression parameters. This procedure uses the approximate 16th, 50th, and 84th percentiles of the X and corresponding Y's. In this paper, we investigate some properties of the estimates of slope and intercept using the median-median line. We extend the idea of using a combination of the percentiles of the X and Y values to estimate the regression parameters. We compare this approach to the least squares procedure and alternative robust regression methods. Finally, we characterize the situations in which this approach exhibits a significant improvement over the least squares line.

email: zubovic@ipfw.edu

65. BAYESIAN METHODS IN CLINICAL TRIALS

PROOF OF CONCEPT TRIALS

A. Lawrence Gould*, Merck Research Laboratories

The term 'Proof of Concept' (POC) is not explicitly defined, although it appears frequently in the literature. PoC trials may be undertaken for many reasons, e.g., screening to identify compounds worth developing or determining the feasibility of potential new uses for approved compounds. Statistical considerations become important for PoC trials when stronger evidence than simply possibility is needed, or when a quantity needs to be estimated with defined variability. This presentation provides a rational definition of PoC trials based on predictive probability of 'success' with the use of an interim evaluation to provide early evidence of clear futility or effect. This definition of PoC trials allows the use of any statistically valid design. For example, demonstrating PoC may require being able to say with 100g% confidence that the true treatment is at least D, with the ability to terminate early if the predictive probability of making this statement is less than pL or greater than pU. The definition requires specifying the degree of uncertainty that is tolerable and the magnitude of the treatment effect to confirm, and presents the issues in terms that management generally uses in approaching a decision.

email: goulda@merck.com

BAYESIAN SAMPLE SIZE CALCULATIONS IN PHASE II CLINICAL TRIALS USING A SIMPLE MIXTURE OF INFORMATIVE PRIORS: INCORPORATING PESSIMISTIC AND OPTIMISTIC OPINIONS SIMULTANEOUSLY

Byron J. Gajewski*, University of Kansas Medical Center
Matthew S. Mayo, University of Kansas Medical Center

A number of researchers have discussed phase II clinical trials from a Bayesian perspective. A recent article by Mayo and Gajewski focuses on sample size calculations, which they determine by specifying an informative prior distribution and then calculating a posterior probability that the true response will exceed a prespecified target. In this article, we extend these sample size calculations to include a simple mixture of informative prior distributions. The mixture comes from two sources of information (clinicians). The first clinician is pessimistic about the drug and the second clinician is optimistic. We tabulate the results for sample size design using the fact that the simple mixture of Betas is a conjugate family for the Binomial model.

email: bgajewski@kumc.edu

BAYESIAN ESTIMATION OF COST-EFFECTIVENESS USING PATTERN-MIXTURE MODELS

Clara Y. Kim*, University of Pennsylvania
Daniel F. Heitjan, University of Pennsylvania

We propose a Bayesian parametric approach to compare the cost-effectiveness of a new treatment to an existing treatment when costs are measured longitudinally and some subjects are censored. We use a pattern-mixture model to describe the joint distribution of the costs and survival. The model assumes that costs, conditional on survival, follow a multivariate normal distribution. We simulate the posterior distribution and deal with censoring via data augmentation. We apply the method to data from a randomized clinical trial of a treatment for a cardiovascular disease. We compare findings with an analysis using Willan and Lin's frequentist nonparametric method.

email: ckim@cceb.upenn.edu

A CARDIOLOGY TRIAL OPTIMAL DESIGN CONSTRAINED BY ETHICAL CONSIDERATIONS

Manuela Buzoianu*, Carnegie Mellon University
Joseph B. Kadane, Carnegie Mellon University

Recently, many experimental design problems have shown difficulties in finding optimal design solution. In a cardiology trial the selection of patients for medical diagnosis is monitored using a Bayesian experimental design. In conducting this experiment we describe the available medical decisions, we gather the experimental data by verifying patient disease status and their characteristics, we perform statistical analysis for these data, and, based on this analysis, we choose the best medical action for any future patient. The optimal design has to be obtained to describe the best set of experimental patients. It is accomplished by maximizing some expected utility. Simulation-based methods are used to evaluate expected utility of each design, and they pose challenging computational problems when optimizing over a discrete high-dimensional design space. Methods based on greedy and stepwise selection idea are developed to overcome such difficulties. After obtaining the optimal design, we see that some of the selected patients get assignments that do not maximize patient's utility. A constraint is imposed on the design to address ethical concerns. We study how much efficacy the trial gives up by limiting assignments to those that maximize patient's utility.

email: manuela@stat.cmu.edu

DO ANTIDEPRESSANTS CAUSE SUICIDE IN CHILDREN? A BAYESIAN META-ANALYSIS

Eloise E. Kaizar*, Carnegie Mellon University
Joel Greenhouse, Carnegie Mellon University
Howard Seltman, Carnegie Mellon University

Determining whether antidepressants, including SSRIs, cause suicide in children and adolescents has been at the forefront of FDA regulatory priorities. To shed light on this question, the FDA collected data from randomized controlled trials in children that included an antidepressant intervention. Among 4,582 young people involved in 24 studies there were no completed suicides. However, the overall rate of suicidal behavior and ideation was 1.7%. The FDA performed a meta-analysis and concluded that the relative risk of suicidal behavior and ideation in antidepressants vs. placebo was 1.95 (1.28, 2.98). The FDA concluded the evidence was sufficiently strong to require a black-box warning on the label of all antidepressants that emphasizes the possible risk of increased suicidality in children and adolescents. This warning is expected to reduce antidepressant prescriptions for children and adolescents. We extend the FDA analysis using Bayesian multi-level models that allow for variability due to different types of antidepressants and different psychiatric diagnoses for which the efficacy of antidepressants was being assessed, e.g., depression and anxiety. We are particularly interested in how statistical issues such as model specification and sensitivity analysis inform the regulatory decision and what we can learn from meta-analysis about the scientific questions of interest.

email: ekaizar@stat.cmu.edu

EXPLORATORY BAYESIAN MODEL SELECTION FOR HIGH-ORDER SNP-PHENOTYPE ASSOCIATIONS

Jing X. Zhao*, Merck Research Laboratories
Andrea S. Foulkes, University of Massachusetts
Edward I. George, University of Pennsylvania
Muredach Reilly, University of Pennsylvania
Daniel J. Rader, University of Pennsylvania

This paper proposes a Bayesian Genotype Selection method to identify multi-locus genetic contributors to complex disease. We demonstrate that this approach is well suited to large model spaces and has reasonable power to detect high order genotype-phenotype associations. Markov Chain Monte Carlo simulations are implemented and the model with the highest visiting frequency is selected as the best model. Permutation tests are conducted to assess the significance of our findings. Data simulations are described to characterize this method's ability to detect true, underlying relationships. An application to patients at risk for cardiovascular disease and 4 single nucleotide polymorphisms in 3 candidate lipase genes is provided.

email: jzhao@cceb.upenn.edu

NORMALIZATION OF MICROARRAYS IN TRANSCRIPTION INHIBITION

Yan Zheng*, University of Minnesota
Cavan Reilly, University of Minnesota

Almost all of the existing methods for normalization assume that not too many of the genes differ in expression levels across arrays. Hence when the mean intensity is not roughly constant across arrays, these standard methods are inappropriate. As an alternative, we present a model which allows a different mean level of gene expression across arrays. Here we develop the model in the context of an experiment that attempts to measure mRNA halflifes by stopping transcription and then measuring gene expression at certain later times. By supposing there are genes with long halflifes relative to the duration of the experiment, the model allows estimation of normalizing terms. Certain weaknesses of the basic model are noted, and a more sophisticated model is developed that includes more of the relevant Biology.

email: yanzheng@biostat.umn.edu

A BAYESIAN METHOD FOR FINDING INTERACTIONS IN GENOMIC STUDIES

Wei Chen*, University of Michigan
Debashis Ghosh, University of Michigan
Trivellore Raghunathan, University of Michigan
Sharon Kardia, University of Michigan

An important step in building a multiple regression model is the selection of predictors. In genomic and epidemiologic studies, datasets with a small sample size and a large number of predictors are common. In such settings, most standard methods for identifying a good subset of predictors are unstable. Furthermore, there is an increasing emphasis towards identification of interactions, which has not been studied much in the statistical literature. In this article, we propose a method, called BSI (Bayesian Selection of Interactions), for selecting predictors in a regression setting when the number of predictors is considerably larger than the sample size with a focus towards selecting interactions. Latent variables are used to infer subset choices based on the posterior distribution. Inference about interactions is implemented by a constraint on the latent variables. The posterior distribution is computed using the Gibbs Sampling methods. The finite-sample properties of the proposed method are assessed by simulation studies. We illustrate the proposed method by analyzing data from a hypertension study involving Single Nucleotide Polymorphisms (SNPs). Keywords: Bayesian variable selection; Conditional prior distribution; Constrained Bayes inference; Gibbs sampling; Latent mixture modeling.

email: lisachen@umich.edu

ESTIMATING MODEL COMPLEXITY FOR BAYESIAN NETWORKS LEARNING

Anthony Almudevar, University of Rochester
Peter Salzman*, University of Rochester

Bayesian networks are commonly used to model complex genetic interaction graphs in which genes are represented by nodes and interactions by directed edges. Although a likelihood function is usually well defined, the maximum likelihood approach favors networks with high model complexity. To overcome this we propose a two step algorithm to learn the network structure. First,we estimate model complexity. This requires finding the MLE conditional on model complexity then using Bayesian updating, resulting in an informative prior density on complexity. This is accomplished using simulated annealing to solve a constrained optimization problem on the graph space. In the second step we use an MCMC algorithm to construct a posterior density of gene graphs which incorporates the informative prior constructed in the first step. Our approach is illustrated by an example.

email: psalzman@bst.rochester.edu

INCORPORATING PRIOR INFORMATION VIA SHRINKAGE: A COMBINED ANALYSIS OF GENOME-WIDE LOCATION DATA AND GENE EXPRESSION DATA

Yang Xie*, University of Minnesota
Wei Pan, University of Minnesota
Keyong S. Jeong, University of Minnesota
Arkady Khodursky, University of Minnesota

Transcriptional control is a critical step in regulation of gene expression. A difficulty arises due to the weak signal and high noise in various sources of data while most current approaches are limited to analysis of a single source of data. A natural alternative is to improve statistical efficiency and power by a combined analysis of multiple sources of data. Here we propose a shrinkage method to combine genome-wide location data and gene expression data to detect the binding sites or target genes of a transcription factor. Specifically, a prior ``non-target'' gene list is generated by analyzing the expression data, and then this information is incorporated into the subsequent binding data analysis via a shrinkage method. There is a Bayesian justification for this shrinkage method. In simulation studies, the proposed method gives higher sensitivity and lower false discovery rate (FDR) in detecting the target genes. In real data example, proposed method can reduce the estimated FDR and increase the power to detect the previously known target genes of a transcription regulator (Lrp) in Escherichia coli. This method can also be used to incorporate other information to microarray data analysis, such as using Gene Ontology (GO) information, to detect differentially expressed genes.

email: yangxie@biostat.umn.edu

EMPIRICAL BAYES INFERENCE FOR PARTIALLY NESTED DESIGNS IN TWO-COLOR MICROARRAY SYSTEMS

Robert J. Tempelman*, Michigan State University

Various empirical Bayes estimation strategies have been proposed for inferring upon differential gene expression between treatments using microarrays. Typically, these methods only consider shrinkage on the lowest hierarchical level, that involving shrinking gene-specific residual variances to an average. For simple designs, this strategy confers greater statistical power for estimating differential expression relative to those methods using only gene-specific inference. However, experimental variability may be defined at higher hierarchical levels in two color microarrays, particularly for designs involving subsampling (i.e. dye-swaps) or split plots (i.e. arrays being the experimental units for one factor and the block for another). This paper extends work originated by Wolfinger and Kass (2000) and more recently by Feng et al. (2004) who demonstrate that the residual likelihood function can be readily written as a product of independent inverted gamma densities in balanced designs. Shrinkage estimation of multiple sources of variability and estimation of the corresponding experimental error degrees of freedom on statistical power are addressed.

email: tempelma@msu.edu

67. METHODOLOGIC ISSUES IN STUDIES INVOLVING CANCER SCREENING

CAUSAL INFERENCE AND SCREENING

James M. Robins*, Harvard University

Screening for cancer can be viewed as a time varying treatment. If the outcome of interest is death from cancer then cancer occurence is a time dependent covariate that is influenced by past treatment (screening) and both influences future treatment (since patients with cancer often have the cancerous organ removed so no further screening of that organ is possible) and predicts later mortality from cancer. The question of public health interest is to determine the optimal screening schedule. I will discuss new statistical methods for estimating the optimal schedule from large HMO data bases of screened subjects.

email: robins@hsph.harvard.edu

BIAS FROM VARIABILITY IN DIAGNOSTIC DELAY

Timothy R. Church*, University of Minnesota

The well known screening biases from lead-time, length-biased sampling, and self-selection can affect the results of observational epidemiologic studies in ways that can be simulated using plausible models that specify distributions for preclinical duration, as well as for incidence and survival. These same models can be applied to the variability between individuals of their delays in seeking diagnosis and care, even for conditions that don't have widely used screening methods. To the extent that these delays are behaviorally determined and correlated with other health behaviors (e.g., poor diet and disregard of symptoms), bias is introduced into the estimation of those behaviors as risk factors for disease. Results from applying such models with plausible parameters either provide assurance or counsel caution in the interpretation of studies targeting risk factors that may be correlated with diagnostic delays. A brief development of the theory and an example will be presented, along with suggested areas of development.

email: trc@cccs.umn.edu

A NOVEL ALTERNATIVE TO CAUSE-SPECIFIC MORTALITY FOR EVALUATING CANCER SCREENING

Marshall M. Joffe*, University of Pennsylvania

Cause-specific mortality is often used for evaluating the efficacy of cancer screening and more generally for evaluating the effect of various exposures or treatments. Cause-specific mortality suffers from two drawbacks as an outcome measure: 1) cause of death is often hard to determine and may be misclassified; and 2) in the presence of death from other causes, cause-specific mortality may be difficult to interpret. I propose a new measure of the effect of screening: its effect for people diagnosed with cancer. I consider how to define and estimate such effects for simple binary diagnosis and consider extensions for time of diagnosis. I discuss difficulties with this approach and illustrate with data from the Health Insurance Plan Study of breast cancer screening.

email: mjoffe@cceb.upenn.edu

68. CURRENT ADVANCES IN MODELING TIME COURSE GENE EXPRESSION DATA

HIDDEN MARKOV MODELS FOR MICROARRAY TIME COURSE AND CELL CYCLE DATA IN MULTIPLE BIOLOGICAL CONDITIONS

Christina Kendziorski*, University of Wisconsin–Madison
Ping Wang, University of Wisconsin–Madison
Ming Yuan, Georgia Institute of Technology

Yuan and Kendziorski recently proposed a method based on hidden Markov models for identifying genes differentially expressed (DE) across multiple biological conditions over time. I will briefly review that approach and discuss extensions that allow for both identification of DE genes and genes affecting cell cycle phase in a study of leukemia.

email: kendzior@biostat.wisc.edu

STATISTICAL METHODS FOR ANALYSIS OF MICROARRAY TIME COURSE GENE EXPRESSION DATA

Hongzhe Li*, University of California, Davis
Fangxin Hong, Salk Institute

Since many biological systems and processes in human health and diseases are dynamic systems, genome-wide time- course gene expression studies can often provide more insights into such systems. Such studies are essential in biomedical research to understand biological phenomena that evolve in a temporal fashion. Rigorous statistical methods for analyzing such time course data are required in order to account for the noisy nature of the microarray-based gene expression data, the potential dependency of the gene expression measurements over time and the time-dependency of the gene expression data. In this talk, I will present some statistical problems and methods for analyzing such time course gene expression data, including the B-spline based hierarchical models and empirical Bayes methods to model gene expression trajectories over time and to detect temporally regulated genes and temporally differentially expressed genes. I will also briefly talk about the model-based method for identifying periodically regulated genes. Both simulated data and real gene expression data sets related to C. elegans developmental process and yeast cell cycle process are used to demonstrate these methods.

email: hli@ucdavis.edu

A SEQUENTIAL BAYESIAN APPROACH WITH APPLICATIONS TO CIRCADIAN RHYTHM MICROARRAY GENE EXPRESSION DATA

Faming Liang*, Texas A&M University
Chuanhai Liu, Texas A&M University
Naisyin Wang, Texas A&M University

This talk focuses on new sequential Bayesian methods motivated by the analysis of circadian rhythm microarray gene expression data collected from avian pineal gland. Data were collected under light-dark (LD) and constant darkness (DD) conditions. Interesting statistical topics, such as identification of differentially expressed genes, clustering of gene expression profiles, and analysis of changes of gene expression profiles under different experimental conditions, will be discussed. We will also present the analysis of avian pineal gland gene expression data.

email: fliang@stat.tamu.edu

69. STATISTICAL ISSUES IN A BIOMETRIC SURVEY: THE NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES)

INDICATORS OF OBESITY FROM NHANES

Allison Hedley*, Centers for Disease Control and Prevention
Cynthia Ogden, Centers for Disease Control and Prevention

The increasing prevalence of obesity in the United States is considered an important public health issue. The National Health and Nutrition Examination Survey (NHANES) is a multistage, national probability sample that provides unique, national level data on the health of the U.S., including indicators of overweight and obesity. NHANES collects information through personal interviews, medical examinations, and laboratory measurements. Standard equipment and protocols are used for each measurement. The definitions of overweight and obesity used by the Centers for Disease Control and Prevention (CDC), the National Institutes of Health (NIH), and the World Health Organization (WHO) rely on the body mass index (BMI, calculated as weight in kilograms divided by the square of height in meters). The national estimates of overweight and obesity in the U.S. are based on BMIs calculated from measured height and weight data collected in NHANES (JAMA, 2004). Other measurements available from NHANES that could be used to evaluate body composition and obesity include self-reported height and weight, bio-electrical impedance (BIA), DXA, measured waist circumference, and skinfold thickness. This presentation will compare the relative strengths and weaknesses of each measure as an indicator of obesity.

email: AHedley@cdc.gov

STATISTICAL ISSUES IN ANALYZING ENVIRONMENTAL HEALTH DATA FROM THE NHANES

Susan Schober*, Centers for Disease Control and Prevention
Lisa Mirel, Centers for Disease Control and Prevention
Lester Curtin, Centers for Disease Control and Prevention

The Centers for Disease Control has recently published the Second National Report on Human Exposure to Environmental Chemicals (CDC, 2003). This report was based entirely on data from the NHANES and includes measures of exposure to dioxins, organophosphate and other contemporary pesticides, polycyclic aromatic hydrocarbons, persistent pesticides, phthalates, heavy metals (including lead and mercury), and tobacco smoke. This presentation will discuss a number of statistical issues that arise in the analysis and interpretation of the environmental data based on a complex probability sample. In particular, estimation of percentile distributions involve additional complications. Many of the environmental measures are done on sub-samples of the data. There may be component non-response related to the various modes of data collection which leads to additional concerns in providing appropriate sample weights for design based analysis. For many chemicals, the proportion of results that fall below the limit of detection is high and in some cases the detection limits are variable, presenting additional statistical complexity. In addition, the statistical issues related to small sample sizes, the number of data years required to provide reliable estimates, and the statistical interpretation of the data will be discussed.

email: SUS2@CDC.GOV

STUDIES OF CARDIOVASCULAR FITNESS IN NHANES

Chia-Yih Wang*, Centers for Disease Control and Prevention
Jeffrey Hughes, Centers for Disease Control and Prevention
Lester Curtin, Centers for Disease Control and Prevention

A cardiovascular fitness component was added to the National health and Examination and Nutrition Survey (NHANES) in 1999. The protocol involves a submaximal exercise test. During the exercise test information is captured on heart rate, systolic blood pressure, and diastolic blood pressure. The test was done only for those 12-49 years of age. For the time period 1999-2002, there were 12,834 persons aged 12-49 sampled, 10,210 (84%) were interviewed and 9,754 (80%) were examined. Of those examined, 69% were tested and 31% were not tested. Those "not tested" include refusals, those who did not have enough time to do the test, and those who met exclusion criteria. There is a lengthy list of criteria for exclusion, including individual health history, safety issues, and use of specified medications. The statistical issue in the analysis of the cardiovascular fitness component becomes one of bias assessment and degree of generalization of results. Specifically, with the level of non response, are the respondents a biased sample and, do the results provide an inference to the entire population or a special subset of the population. The bias assessment for the 1999-2202 data is presented along with a discussion of the methodological issues pertinent to the analysis of this data.

email: lrc2@cdc.gov

---

70. STRUCTURAL EQUATIONS AND PSYCHOMETRIC METHODS IN BIOLOGICAL STUDIES

STRUCTURAL MODELS WITH NONNORMAL, MISSING, MULTILEVEL, OR ATYPICAL DATA: THE EQS APPROACH

Peter M. Bentler*, University of California, Los Angeles

In the Bentler-Weeks model, the parameters of any linear structural equation model [regression, multivariate regression, confirmatory factor analysis, growth curves, lisrel-type model, etc.] are the regression coefficients and the variances and covariances of independent (non- dependent) variables, and possibly the means of independent variables. These provide the mean structure (if any) and covariance structure of the model. The statistical theory for estimation and testing of mean and covariance structures is reviewed, and applied to several types of data structures. This theory is translated into an equation by equation specification in Multivariate Software's EQS Structural Equations Program. The 7 equation example model below has 6(7)/2=21 sample covariances and 13 free parameters, hence 8 df. Latent variable F1 (with indicators V1-V3) is regressed on latent variable F2 (with indicators V4-V6). Free parameters are indicated by "*". /SPECIFICATIONS VARIABLES=6; CASES=200; METHOD=ML,ROBUST; !if nonnormal /EQUATIONS !for 7 dependent variables V1= F1+E1; V2=*F1+E2; V3=*F1+E3; !F1 measurement model V4= F2+E4; V5=*F2+E5; V6=*F2+E6; !F2 measurement model F1 = *F2 + D1; !regression of F1 on F2, with residual D1 /VARIANCES !of 8 independent variables D1=*; F2=*;E1 TO E6=*; /COVARIANCES !of independent variables (none here) /MEANS !not needed (no mean structure) /END

email: bentler@ucla.edu

SPARSE REGRESSIONS AND GRAPHICAL MODELS

Mike West*, Duke University
Adrian Dobra, Duke University
Chris Hans, Duke University
Carlos Carvalho, Duke University

Increasing opportunity to make observation on high-dimensional variables, such as high-throughput molecular data in genomics, raises challenges to statistical science as parameter dimension scales drastically. One key concept that seems to be fundamental to scaling methodology is sparsity. Our work with large-scale regressions and graphical models provides a number of examples of how coherent Bayesian models can be developed and applied in problems of very high dimensions, underpinned by the emphasis through sparsity inducing priors on sparse structure in multivariate relationships that is consistent with both general scientific parsimony as well as the substantive context of motivating applications in exploratory gene expression analysis. Questions of model search are addressed through distributed computational approaches such as our shotgun stochastic search for rapid identification and evaluation of millions of billions of models. This talk will review and discuss aspects of this work, with some examples, including visualization tools for model and inferential exploration, and will highlight open, challenging questions we all face in dealing with model uncertainty in large spaces of statistical models.

email: mw@isds.duke.edu

BAYESIAN LATENT VARIABLE DENSITY REGRESSION

David B. Dunson*, National Institute of Environmental Health Sciences

In hierarchical models with latent variables, there is commonly uncertainty in the latent variable and/or outcome distributions, and these distributions can potentially change with predictors. For example, in studies of DNA damage and repair, the comet assay is used to obtain multiple surrogates of the frequency of DNA strand breaks for individual cells. The distributions of these surrogates do not follow standard parametric forms and can change in shape with genetic and environmental factors. This talk proposes semiparametric Bayesian methods for density regression based on a novel nonparametric approach, which allows a random probability distribution to change flexibly with multiple discrete and continuous predictors. The approach has a number of appealing practical and theoretical properties, which are illustrated through application to epidemiologic data.

email: dunson1@niehs.nih.gov

MIXED-EFFECTS VARIANCE COMPONENTS MODELS FOR BIOMETRIC FAMILY AND LONGITUDINAL ANALYSES

John J. McArdle*, University of Virginia
Carol A. Prescott, Virginia Commonwealth University

Recent research employing biometric analyses of twin and family data has highlighted the benefits of structural equation models (SEM) and mixed effect multilevel (MEML) modeling techniques. We first highlight the exact algebraic equivalence between the standard biometric path analysis (PA) and the corresponding variance components (VC) approaches to modeling family data. Second, we demonstrate how several SEM programs based on the PA and VC approach produce equivalent estimates for all phenotypic and biometric parameters. We then show how the VC approach can be easily programmed using mixed-effect or multi-level model (MEML) programs (e.g., SAS PROC MIXED). We then expand these models to include measured covariates, observed variable interactions, multivariate longitudinal data, and multiple relatives within each family. Software to fit mixed effect models may have advantages over SEM software for programming complex models, including the flexibility of data input, treatment of missing data, inclusion of covariates, and ease of accommodating varying numbers of observations (per family or individual).

---

71. SEEKING PATTERN IN GENOMICS DATA USING CROSS-SPECIES COMPARISONS

GLOBAL CLASSIFICATION OF (PLANT) PROTEINS ACROSS MULTIPLE SPECIES

Kerr Wall, Pennsylvania State University
Jim Leebens-Mack, Pennsylvania State University
Naomi S. Altman*, Pennsylvania State University
Claude dePamphilis, Pennsylvania State University
Victor Albert, Natural History Museums and Botanical Garden, University of Oslo
Dawn Field, Oxford University

With rapidly growing numbers of whole genome and expressed sequence tag (EST) sequences in our public databases, sequence-based protein classification systems are providing foundations for gene annotation, functional genomics, and comparative investigations of gene and genome evolution. We use the similarity-based clustering procedure TribeMCL (Enright et al 2002, 2003) to classify protein-coding genes into putative gene families for rice and protein. The results of these analyses provide insights into the Arabidopsis and rice genomes, gene family evolution, and the evolutionary dynamics of functional domains among gene families. Phylogenetic analyses of exemplar gene families shows a strong, but not perfect correspondence between tribe membership and cladistic relationships. One of the challenges of this type of classification is determining the stability of the proposed families under small changes in the data. Classifications have been constructed using three clustering stringencies and the strength of support for clusters as historical entities has been tested through jackknife analyses and “bagging” methodology.

email: naomi@stat.psu.edu

STUDYING FUNCTIONAL NON-CODING SEQUENCES THROUGH SUPERVISED AND UNSUPERVISED ANALYSES OF GENOMIC ALIGNMENT DATA

James Taylor, Pennsylvania State University
Webb Miller, Pennsylvania State University
Francesca Chiaromonte*, Pennsylvania State University

According to recent estimates, as much as 3-4% of the human genome may be involved in functions other than coding for proteins. The study of functional non-coding sequences is an emerging focus of genomics and bioinformatics, with comparative approaches made possible by the availability of whole-genome alignments of several species. In this talk, we describe supervised and unsupervised analyses that exploit short alignment pattern information to identify and characterize putative functional sites. As an instance of supervised analyses, we present Regulatory Potential (RP) scores, which predict regulatory elements based on alignment patterns between human and rodents. As an instance of unsupervised analyses, we present preliminary results on the clustering of non-coding sequences based on alignment patterns between human and chicken.

email: chiaro@stat.psu.edu

INVESTIGATION OF PLANT PHOSPHORYLATION USING INTERGENOMIC COMPARISON

Michael Gribskov*, Purdue University

Signal transduction refers to the process by which external stimuli such as temperature, nutrients, environmental stress, and pathogens, effect cellular growth and development. Much of this process is mediated by cascades of phosphorylation in which protein kinases phosphorylate a target, the target (often a protein kinase) is activated by the phosphorylation, and in turn phosphorylates other proteins. Phosphorylation is thus a switch allowing cells to make very rapid and complex responses to external signals. Plants pose a particularly interesting problem in understanding signal transduction. Not only do plants possess many more (about three times) protein kinases than animals, but far fewer of them have received detailed biochemical study. In this presentation I will discuss how comparative genomics can be used to make inferences about the functions, targets, and interactions of plant protein kinases.

email: gribskov@purdue.edu

CONFIDENCE INTERVALS FOR SEMI-PARAMETRIC QUANTILE REGRESSION

Mi-Ok Kim*, University of Kentucky

Quantile regression offers an alternative to least squares by providing a more complete picture of the response conditional distribution. The regression has been explored with semi- parametric models by He and Shi (1998), He and Liang (2000), and Lee (2003). The studies discuss the estimation of the regression and related theoretical properties, leaving the inference unstudied. We focus on confidence intervals for the linear part in the model. Several methods, some based on resampling and some on the asymptotic distribution of the estimates are introduced. While Wald- type confidence intervals based on the asymptotic distribution of the estimates seem a natural choice, it presents a challenge of estimating the error density at the conditional quantile of interest (well known in a linear model). On the other hand, resampling based confidence intervals offer viable alternatives. We explore different resampling techniques such as estimating function bootstrap by Hu and Kalbfleisch (200), perturbing the minimand by Jin, Ying and Wei (2001), and Markov Chain Marginal Bootstrap (MCMB) by He and Hu (2002). We conduct a Monte Carlo study to compare their performances.

email: miokkim@uky.edu

CAPTURING HIGHER-ORDER FEATURES IN POOLING STRATEGY WITH KERNEL LOGISTIC MODEL

Peter X. K. Song, University of Waterloo
Peng Zhang*, University of Waterloo
Rui Liu, York University

It is often expensive or impossible in practice to get individual measurement due to budget or other limitations. Pooling strategy was proposed for case-control study to reduce the number of assays with the entire sample. However, it is also reported that such approach may fail when interaction or higher order terms included in the logistic model. We propose a nonparametric approach to handle data with higher order feature, which is to capture the curvature pattern of the data using kernel logistic model with only linear terms. Through simulation study we illustrate and compare the effects of this method with different pooling size.

email: p5zhang@math.uwaterloo.ca

SEMIPARAMETRIC REGRESSION FOR HIGH-DIMENSIONAL DATA WITH APPLICATIONS IN MICROARRAYS: LEAST-SQUARE KERNEL MACHINES AND LINEAR MIXED MODELS

Dawei Liu*, University of Michigan
Xihong Lin, University of Michigan
Debashis Ghosh, University of Michigan

We consider a semiparametric regression model for modelling high dimensional covariate data, e.g., microarrays. This model relates a normal clinical outcome to clinical covariates and gene expressions, where the clinical covariate effects are modelled parametrically and gene expression effects are modelled nonparametrically using least square kernel machines (LSKMs). The nonparametric function of gene expressions allows for the possibility that the number of genes might be large and the genes are likely to interact with each other. We show that the dual problem derived from the primal problem of the least square kernel machine can be formulated using a linear mixed effects model. Estimation hence can proceed within the linear mixed model framework using standard mixed model software. Both the regression coefficients of the clinical covariate effects and the least square kernel machine estimator of the nonparametric gene expression function can be obtained using the Best Linear Unbiased Predictor in linear mixed models. The smoothing parameter and the kernel scale parameter can be estimated as variance components using REML in linear mixed models. A bootstrap test is developed to test for significant gene expression effects. The methods are illustrated using a prostate cancer data set and evaluated using simulations.

email: liudawei@umich.edu

ON A FLEXIBLE INFORMATION CRITERION FOR ORDER SELECTION IN FINITE MIXTURE MODELS

Richard Charnigo*, University of Kentucky
Ramani S. Pilla, Case Western Reserve University

Finite mixture models provide easily-interpreted representations of the heterogeneity in physical phenomena and biological processes; yet, finite mixture models pose special challenges to statisticians, especially with regard to estimation of the order (i.e., the number of distinct mixture components). Lindsay (1983) has developed an elegant framework for nonparametric estimation of the mixing distribution (and, hence, of the order) in the absence of a structural parameter common to all mixture components. However, we demonstrate that, under fairly general conditions, incorporation of a structural parameter results in nonexistence of the semiparametric estimate or in a degenerate semiparametric estimate. Thus, a different paradigm for order selection is required to accommodate the presence of a structural parameter. We propose a flexible information criterion (FLIC) by which both the order of a finite mixture model and the value of the structural parameter can be consistently estimated. The FLIC is similar in spirit to the AIC and BIC but is adaptive in the sense that the strength of the penalty is determined by the data, a feature absent from the AIC and BIC. We investigate the performance of the FLIC through simulation experiments and applications to real data sets.

email: richc@ms.uky.edu

TESTING LACK-OF-FIT OF NONLINEAR REGRESSION MODELS VIA LOCAL LINEAR REGRESSION TECHNIQUES

Chin-Shang Li*, St. Jude Children's Research Hospital

A data-driven test is proposed to assess the lack of fit of nonlinear regression models. This test is based on comparison of local linear and parametric fits, and no boundary-corrected kernels are needed at the boundary when local linear fitting is used. The asymptotically optimal bandwidth can be used for bandwidth selection under the parametric null model. This selection method leads to the data-driven test that has a limiting normal distribution under the null hypothesis and is consistent against any fixed alternative. The finite-sample property of the proposed data-driven test is illustrated, and the power of the test is compared with some existing tests through simulation studies. A real-life data set is used to illustrate the practical use of the proposed test.

email: chinshang.li@stjude.org

NONPARAMETRIC SPLINE ESTIMATORS OF COMPARISON DISTRIBUTION FUNCTIONS AND ROC CURVES

Piea Peng Lee*, Macquarie University, Sydney, Australia
Andrzej Kozek, Macquarie University, Sydney, Australia

We consider splines with simple, equally spaced knots for estimating the comparison distribution functions of two populations. The distributions of the two populations are investigated for homogeneity. Our spline estimators of the comparison distribution functions are constructed by means of linear local positive spline operators. We determine the optimal uniform spacings for the knots of the splines empirically. We show that our estimators are uniformly consistent and we derive their asymptotic distribution. Simulations show that our spline estimators of comparison distribution functions work well. We are also applying the spline approximation technique to the receiver operating characteristic (ROC) curves. The latter are used as graphical tools in medical research for drawing a distinction between diseased and healthy individuals.

email: plee@efs.mq.edu.au

IMPROVING REGRESSION FUNCTION ESTIMATORS

Ali Khoujmane*, Texas Tech University

This research is concerned with estimation problems regarding nonparametric regression functions that are not necessarily directly observed. Practical examples are abundant; for instance one may want to infer about the weight distribution of a cable of which only the shape is known. In general one may think of an input-output system, where one wants to recover an unknown parameter of the input. At least two measurement designs can be employed: random design, where the points at which the output function is observed are chosen according to a random mechanism; or deterministic design where these points are chosen essentially error free by the observer (for instance equally distant points in the unit interval). The random design model leads statistically to independent and identically distributed observations. This is no longer true for the deterministic design where the data are independent but not identically distributed. Therefore the latter situation is mathematically somewhat harder to deal with than the former, and most of the results in the literature, whenever available at all in this rather complicated model, are usually formulated and proved for the independent and identically distributed case.

email: akhoujma@math.ttu.edu

73. IMAGING OF BRAIN ACTIVITY

INCREASING THE POWER OF GROUP COMPARISONS IN SPECT BRAIN IMAGING THROUGH SPATIAL MODELING OF INTERVOXEL CORRELATIONS

Jeffrey S. Spence*, University of Texas Southwestern Medical Center
Patrick S. Carmack, University of Texas Southwestern Medical Center
Robert W. Haley, University of Texas Southwestern Medical Center
Richard F. Gunst, Southern Methodist University
William R. Schucany, Southern Methodist University
Wayne A. Woodward, Southern Methodist University

The size and complexity of brain imaging databases confront statistical analysts with many issues when attempting to determine brain activation differences between groups of subjects. Among these are statistical issues, such as the analysis of hundreds of thousands of spatially correlated measurements per image. In this article, an analysis of SPECT brain images is used to detail how power in group comparisons can be increased. Key to increasing power is the spatial modeling of strong intervoxel correlations. Exploiting this correlation, blocks of contiguous voxels are defined within several regions of the deep brain. Using kriging methods, block averages and their prediction variances are calculated, and these provide the data for a general linear model (GLM) analysis of group effects. This GLM analysis is shown to be much more powerful than the voxel-by-voxel analysis commonly used by medical researchers. These procedures are applied to comparisons of SPECT brain-imaging data from four groups of subjects, three of which are identified as having variants of the 1991 Gulf War syndrome and one of which is a control group. Spatial modeling and analyses of these data identify structures in the basal ganglia and the brain stem that exhibit statistically significant group differences in cholinergic response to a physostigmine challenge.

email: Jeffrey.Spence@UTSouthwestern.edu

A HIERARCHICAL DEFORMATION MODEL FOR IMAGES

Sining Chen*, Sidney Kimmel Comprehensive Cancer Center
Helene Benveniste, Brookhaven National Laboratory
Valen E. Johnson, University of Texas M. D. Anderson Cancer Center

Large quantities of medical images are acquired daily at nearly every medical center in the United States, but statistical models and associated software to facilitate automated analyses of these images is lacking. In this paper we develop methodology aimed at filling at least part of this void. The proposed methodology is based on an atlas-based deformation model which identifies a one-to-one mapping from the atlas image to a target image within the same image class. The deformation is based on co-registering a set of generalized landmarks points which we call ``facets''. The model for facet locations has two components: a Markov random field prior defined with respect to a nearest-neighbor clique system, and a data component that measures the agreement of features at registered landmark points in the atlas and target image. Using a new feature similarity measure, we find that the model performs approximately as well as a human operator for a particular class of image segmentation tasks.

email: schen46@jhmi.edu

A BAYESIAN APPROACH TO DETERMINING CONNECTIVITY OF THE HUMAN BRAIN

Rajan S. Patel*, Emory University
F. Dubois Bowman, Emory University

We develop a novel descriptive and inferential method to analyze and compare the connectivity of the human brain within and between subject groups using functional magnetic resonance imaging (fMRI) data. We assess the relationship between pairs of distinct brain regions by comparing expected joint and marginal activation probabilities of voxel pairs. We utilize a Bayesian paradigm, which allows us to incorporate known anatomical and functional information to determine these probabilities. We define the relationship between two distinct brain regions by distinct measures of functional connectivity and influence. After assessing the relationship between all pairs of brain voxels, we are able to construct functional networks from any given brain region and assess significant functional connectivity and influence in these networks. Our method can be used to develop causal brain networks for use with structural equation modelling. Our approach has important practical implications in practice by allowing an investigation into the differences in connectivity between two groups, such as a patient groups diagnosed with a psychiatric disorder and normal control subjects. We illustrate the use of our connectivity analysis using fMRI data from a study of subjects while playing the Prisoner's Dilemma game.

email: rspate2@emory.edu

A MIXTURE APPROACH FOR PET STUDIES

Huiping Jiang*, New York State Psychiatric Institute
Todd Ogden, Columbia University

A new kinetic modeling approach for quantification of dynamic position emission tomography (PET) studies in humans is presented. The approach expresses the time-course data as a weighted linear combination of a finite number of basic functions (components) given by a sum of exponential functions convoluted with an arterial blood input function. The selection of the number of components is based on AIC, and the parameter estimates can be obtained using a nonlinear weighted least squares method. The approach can be applied both in a voxel-based analysis and a region of interest (ROI) analysis. In addition, the proposed approach has faster processing time than others. The method is illustrated with analysis of both simulated and real data.

email: jianghu@pi.cpmc.columbia.edu

3D WAVELET DENOISING OF SPECT IMAGES

Liansheng Tang*, Southern Methodist University
William R. Schucany, Southern Methodist University
Wayne A. Woodward, Southern Methodist University

The quality of SPECT (single photon emission computed tomography) three-dimensional (3D) human brain images is adversely affected by the noise inherent in photon counts. To improve the signal-to-noise ratio, 3D wavelet denoising may be preferred to Gaussian smoothing. A discrete wavelet transform via Mallat's pyramid algorithm is performed on the original SPECT image and denoising and enhancement are obtained by thresholding of the wavelet coefficients in the transform domain using adaptive shrinkage (Donoho-Johnstone 1994). The inverse wavelet transform of the thresholded coefficients is used to reconstruct the images. We compare these results with those of Gaussian-smoothed images for detecting areas of activation using the SPM brain-imaging software.

email: ltang@mail.smu.edu

THE BRAIN AS A MEDIATOR: WHERE DOES THIS AFFECT THAT?

David M. Shera*, University of Pennsylvania
Lijun Jing, University of Pennsylvania
Tony J. Simon, University of Pennsylvania

We have implemented a model to look at structural MRI brain image measures as mediators of a relationship between a genetic difference and specific neurocognitive differences. The classic mediation model matches the biology particularly well and we have implemented it to examine MRI brain images on a voxel-wise basis. The candidate mediator is highly multivariate and results can be partitioned into four distinct groups, two of which are consistent with the mediation model and two of which are inconsistent. Inferences based on the different groups can lead to determining structure/function relationships. We apply the method to data from an investigation of Chromosome 22q11.2 Deletion Syndrome and specific cognitive deficits. Our results corroborate those of other analysis methods.

email: shera@email.chop.edu

74. COX REGRESSION MODELS

BAYESIAN VARIABLE SELECTION IN COX REGRESSION MODELS

Naijun Sha*, University of Texas at El Paso
Mahlet Tadesse, University of Pennsylvania
Marina Vannucci, Texas A&M University

In this paper, we investigate variable selection methods for Cox's proportional hazard model. We develop selection methods that allow for censored data. Our methods lead simultaneously to an estimate of the survival function as well as to the identification of the factors that affect the survival outcome. We handle the problem of selecting a few predictors among the prohibitively vast number of variables through the introduction of a binary exclusion/inclusion latent vector. This vector is updated via an MCMC technique to identify promising models. We describe strategies for posterior inference and explore the performance of the methodology with simulated and real datasets.

email: naijun@math.utep.edu

GENERAL INSTRUMENTAL VARIABLES ESTIMATION IN COX'S PROPORTIONAL HAZARDS MODEL WITH TIME-VARYING TREATMENT

David S. Cohen*, Johns Hopkins University
Thomas A. Louis, Johns Hopkins University
Daniel O. Scharfstein, Johns Hopkins University
Sam Bozzette, Veterans Affairs San Diego Healthcare System
Henry K. Tam, Veterans Affairs San Diego Healthcare System
Christopher F. Ake, Veterans Affairs San Diego Healthcare System

We propose the use of a general instrumental variable approach to Cox's proportional hazard model with time-varying treatment to account for possible selection bias and unmeasured confounders. Within each small interval of time, we exploit the properties of an instrumental variable to decompose the probability of survival given the instrument into a structural component relating failure to treatment and a treatment selection component relating treatment to instrument. The necessary assumptions and causal interpretation of the estimator are discussed. We apply our method to a large subsample of a Veterans' Affairs database of healthcare services provided to HIV positive men with cardiovascular events as the primary outcome. We take as treatment the cumulative monthly exposure to combination anti-retroviral therapy and as instrument the site at which treatment was prescribed. We compare results to a covariate adjusted approach.

email: dcohen@jhsph.edu

ASSESSMENT OF THE COX MODEL FOR BINARY INTERVAL-CENSORED FAILURE TIME DATA

Lianming Wang*, University of Missouri

This paper discusses statistical analysis of binary interval-censored failure time data with focus on the marginal Cox model approach. The approach assumes that the marginal distributions for the correlated failure times can be described by the Cox model and leaves the dependence structure completely unspecified. It is apparent that one important question for the approach is to assess the model adequacy since the statistical inference depends critically on the veracity of the assumed model. For correlated right-censored failure time data, some methods have been proposed for checking the appropriateness of the marginal Cox model (Spiekerman and Lin,1996). This paper considers correlated interval- censored data,for which there is no method available for model checking, and a goodness-of-fit test is proposed for the problem. The method is applied to a set of binary interval-censored data arising from an AIDS clinical trial.

email: lwdzc@mizzou.edu

INCORPORATING TIME-DEPENDENT COVARIATES IN SURVIVAL ANALYSIS USING THE LVAR METHOD

Yali Liu*, Purdue University
Bruce A. Craig, Purdue University

In survival analysis, use of the Cox proportional hazards model requires knowledge of all covariates under consideration at every failure time. Since failure times rarely coincide with observation times, time-dependent covariates (covariates that vary over time) need to be inferred from the observed values. In this paper, we introduce the last value auto-regressed (LVAR) estimation method, demonstrate, through a simulation study, that this method results in a smaller mean square error of the covariate effect compared to others in the literature, and apply the method to a real problem involving Primary Biliary Cirrhosis data from the Mayo clinic. The application shows that LVAR results in stronger effects of log albumin and log prothromin time than several published methods.

email: ylliu@stat.purdue.edu

A COMPARISON OF STATISTICAL TESTS FOR ASSESSING THE PROPORTIONAL HAZARDS ASSUMPTION IN THE COX MODEL

Inger Persson, Trial Form Support, AB, Stockholm, Sweden
Harry J. Khamis*, Wright State University

A comparison of some of the most common numerical procedures to check the assumption of proportional hazards for the Cox model is presented. In a simulation study, fifteen test statistics are evaluated under proportional hazards and five forms of nonproportional hazards: increasing, decreasing, crossing, diverging, and nonmonotonic hazards. The tests are compared in the two-sample case and for Type I censoring. Results indicate that the Gill and Schumacher test has relatively high power for crossing hazards, and the time-dependent covariate test by Cox with time function exp(t) has relatively high power for nonmonotonic hazards. Also, the Breslow, Edler and Berger test with rank scores, the Gill and Schumacher test, and the Harrell test have one of the top four powers for three or more of the five forms of nonproportional hazards.

email: harry.khamis@wright.edu

EXAMINING MODEL FIT FOR EXPOSURE-RESPONSE CURVES WITH PENALIZED SPLINES IN COX MODELS

Elizabeth J. Malloy*, Harvard University
Ellen A. Eisen, Harvard University

Nonparametric models are often used in occupational epidemiology to capture nonlinearities in exposure-response curves. We examined alternate measures of model fit for penalized splines in the Cox proportional hazards model. Relative risk of rectal cancer mortality was modeled in a cohort of 46,400 autoworkers exposed to metalworking fluids. Cox models were fit using a penalized spline for cumulative exposure to metalworking fluids. We examined the sensitivity of the model to degrees of freedom, number of knots, data transformations, and outlier deletion. Model fit criteria, such as AIC, corrected AIC, and cross-validation, were used to determine a limited set of best models that summarized the relationship between cumulative exposure and risk. Shapes of the exposure-response curves that optimized the model fit criteria were similar in the lower exposure range, where the data were most dense. In that region, the curves showed the same increasing relative risk of death from rectal cancer with increasing exposure. In contrast, in the sparser data region, where exposure was higher, the curves varied in shape. Here relative risk declined, remained flat or increased slightly with increasing exposure. Results suggest the distribution of exposure plays an important role in the sensitivity of penalized splines and model fit.

email: emalloy@hsph.harvard.edu

TESTING THE PROPORTIONAL ODDS MODEL FOR INTERVAL-CENSORED SURVIVAL DATA

Jianguo Sun, University of Missouri–Columbia
Liuquan Sun, Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing
Chao Zhu*, University of Missouri–Columbia

This talk discusses the two sample goodness-of-fit test for the proportional odds model based on interval-censored failure time data, which frequently occur in, for example, periodic follow-up survival studies. The proportional odds model has a feature that allows the ratio of two hazard functions to be monotonic and converge to one and provides an important tool for the modeling of survival data. To test the model, we generalize the method given in Dauxois and Kirmani (2003) for right-censored data. The asymptotic distribution of the generalized procedure is established and its finite sample properties are evaluated by simulation studies. For illustration, we apply the proposed test to some well-known interval-censored datasets.

email: cznm4@mizzou.edu

MARGINALIZED TRANSITION MODELS FOR LONGITUDINAL POLYTOMOUS DATA

Keunbaik Lee*, University of Florida
Michael Daniels, University of Florida

Generalized linear models with serial dependence are commonly used for longitudinal data analysis. Heagerty(2002) has suggested marginalized transition models for the analysis of longitudinal binary data. In this paper, we extend his work to accommodate longitudinal ordinal data using cumulative logit model. Fisher-scoring and EM algorithms are developed for estimation. Methods are illustrated on quality of life data from a colorectal cancer clinical trial.

email: lee@stat.ufl.edu

JOINT ANALYSIS OF LONGITUDINAL DATA WITH INFORMATIVE RIGHT CENSORING

Mengling Liu*, New York University School of Medicine
Zhiliang Ying, Columbia University

In this paper, we propose the joint model for the longitudinal data along with the presence of the informative censoring times. A mixed-effects model is used for the longitudinal process and a semiparametric transformation model is proposed for the censoring times. Both model assumptions are broadly used in their own subject areas and linked by a common random effect which is introduced to accommodate the correlation among within- subject responses and dependence between the censoring times and the longitudinal observations. The joint model can correct the bias caused by the informative censoring effectively and a class of consistent and asymptotically normal estimators is developed to estimate the parameters of interest. Moreover, the asymptotic variance of the proposed estimator has the closed form and is readily obtained by plug-in rules. The method is illustrated by simulation and is applied to the Renal Disease data.

email: mengling.liu@med.nyu.edu

USING TRAJECTORIES FROM A BIVARIATE GROWTH CURVE OF COVARIATES IN A COX MODEL ANALYSIS

Qianyu Dang*, University of Pittsburgh
Sati Mazumdar, University of Pittsburgh
Stewart J. Anderson, University of Pittsburgh

In many psychiatric clinical trials of maintenance treatment, patients are first admitted into an open treatment period. If they respond, then they enter the second phase of the trial where they are randomized to the different arms of the "maintenance treatments". Often, more than one variable is measured longitudinally in the open phase of the trial to measure treatment response. Their trajectories of responses to the treatment during the open period are important factors in the later phase of the trial. Using the estimated trajectories of each subject from a bivariate growth curve as predictors, we developed a method for a Cox regression model. To adjust for the estimation errors of the predictors, we applied a full likelihood approach based on the conditional expectation of the covariates. Simulation studies indicate that the error corrected estimators for the model parameters are generally less biased when compared to the naïve regression estimators without accounting for estimation errors. An illustrative example is provided with data from a maintenance treatment trial for major depression in an elderly population. Visual Fortran 90 and SAS IML programs were developed to implement the algorithm.

email: dangq@upmc.edu

FUNCTIONAL DATA ANALYSIS ISSUES FOR IDENTIFYING PLACEBO RESPONSE IN DRUG TREATED SUBJECTS

Thaddeus Tarpey*, Wright State University
Eva Petkova, Columbia University
Todd Ogden, Columbia University

Differentiating placebo response from a true drug response in longitudinal depression studies gives rise to a variety of problems associated with functional data analysis. For instance, suppose the active treatment arm of the study can be modeled by two latent classes corresponding to responders and non-responders. These two latent classes may each consist of a variety of distinct functional response profiles. We propose a mixture modeling approach followed by a prototype analysis (e.g. k-means, principal points). The mixture analysis will determine the different latent classes and the prototype analysis will determine the variety of different response profiles in each latent class. Another issue to consider is the selection of basis functions to represent the functional data. We shall investigate the use of customized basis functions using a Gram-Schmidt procedure that will identify primary modes of variation in response profiles in longitudinal studies.

email: thaddeus.tarpey@wright.edu

ESTIMATING HEALTH OUTCOMES TRAJECTORIES VIA FINITE MIXTURE MODELS

Jason T. Connor*, Carnegie Mellon University
Susana Arrigain, Cleveland Clinic Foundation

We use a publicly available SAS procedure, PROC TRAJ (www.ncovr.org), to estimate longitudinal finite mixture models with a range of link functions. Health outcomes trajectories are polynomial functions across time and individuals. Both a set of unique trajectories and each individual's probability of membership in each trajectory are estimated. We estimate group trajectories of SF-36 scores (normal), probability of alcohol recidivism (binomial), and drinks per day (Poisson) over the course of follow-up for 136 patients in a multi-center liver disease study. Furthermore we estimate the effect of time-varying covariates (e.g. alcohol recidivism) on SF-36 scores. For both SF-36 scales 5 groups were estimated, 4 flat trajectories spaced high to low with one additional group with a rapidly decreasing PCS, and one group that sees steady improvement in MCS scores throughout follow-up. Using time-varying covariates for drinking, drinking is associated with decreases in MCS scores for the two middle function groups. Three unique trajectory groups are clearly identified for redrinking and drinks per day: group 1 remains alcohol free, group 2 begins drinking again immediately and drinks, on average, 2 drinks per day, and group 3 remains alcohol free for 2 years then increases to 1.5 drinks per day at long term follow-up.

email: jconnor@stat.cmu.edu

A LIKELIHOOD-BASED APPROACH TO THE ANALYSIS OF COLONIC CRYPT SIGNALING

Kimberly L. Drews*, Texas A&M University
Raymond J. Carroll, Texas A&M University

There has long been speculation that nearby colonic crypts exhibited coordinated responses in biologically significant events. Our work examines this hypothesis. The responses have a natural hierarchical structure which we posit follows a mixed model, which is difficult to analyze with traditional methods due to the fact that the set of distances between colonic crypts for each subject is unique. We propose the crypt correlations follow AR(1) or Màtern correlation models and note that the interest is in the lowest level of the hierarchy rather than the top level. For data of this type we have developed methodology allowing us to avoid fitting the entire hierarchical model thus permitting us to efficiently find maximum likelihood estimates for the value of the correlation. The method is illustrated using simulations. We then apply our method to a real data set exploring if a coordinated response exists in regards to p27, a cyclin dependent kinase inhibitor which regulates the progression of the cell out of the G1 phase into the S phase and which it is believed helps promote apoptosis.

email: kdrews@stat.tamu.edu

CASE-CONTROL STUDIES WITH LONGITUDINAL COVARIATES

Honghong Zhou*, University of Michigan
Xihong Lin, University of Michigan
Bin Nan, University of Michigan

We consider case-control studies with longitudinal covariates. In a case-control study, subjects are recruited according to their case and control status. In the presence of longitudinal covariates, the covariates are collected longituinally but retrospectively. We assume the longitudinal covariates follow a linear mixed model and the promary binary outcome relates to longitudinal covariates through latent subject specific random intercepts and slopes. We show that the primary binary outcome and the longitudinal covariate can be analyzed jointly through maximum likelihood under rare disease assumption. We apply the method to a case-control study on breast cancer and simulation studies are conducted to evaluate the performance of the method.

email: zhouh@umich.edu

A STRATEGY FOR ASSESSING THE PERFORMANCE OF ALTERNATIVES TO CD4 CELL COUNT IN MAKING THE DECISION TO START ANTIRETROVIRAL TREATMENT IN RESOURCE CONSTRAINED SETTINGS

Lisa M. Wruck*, New England Research Institutes
Michael D. Hughes, Harvard University

To determine timing of initiation of antiretroviral therapy in patients with HIV infection, it is common practice to monitor a marker such as CD4 count repeatedly over time. Patients are recommended for starting treatment once the marker has reached a defined threshold. However, monitoring of CD4 count is difficult in resource constrained settings. We consider the problem of evaluating alternative markers to CD4 count for determining when to start antiretroviral therapy in these settings. To compare treatment start times based on various, potentially complex strategies, we propose a simulation based approach. The trajectory of the marker in question is modeled jointly with CD4 count trajectory and then trajectories are simulated for a very large cohort. A candidate strategy is then applied to the simulated cohort. We use data from the Multicenter AIDS Cohort Study to examine the performance of total lymphocyte count (TLC) as an alternative to CD4 count. We consider cutpoint selection and examine marginal distributions of treatment start times, the distribution of differences in treatment start times within an individual, and distribution of CD4 count at the time TLC crosses its threshold.

email: lwruck@neri.org

SAMPLE SIZE DETERMINATION WITH FDR ADJUSTMENT FOR MICROARRAY EXPERIMENTS

Gengqian Cai*, Temple University
Sanat K. Sarkar, Temple University

False discovery rate (FDR) has been widely used in exploratory research where multiple hypotheses are tested simultaneously, such as in microarray analysis. However, not much progress has been made on the issue of sample size calculation with FDR controlling requirement. This paper investigates the powers and addresses the related problem of sample size determination for current FDR controlling procedures under a mixture model involving independent normal distributions. A general method for sample size calculation is proposed for large number of hypotheses.

email: gcai@temple.edu

MULTIVARIABLE BINARY REGRESSION WITH STOCHASTIC COVARIATES

Evrim Oral*, Middle East Technical University
Diana L. Miglioretti, Center for Health Studies, Group Health Cooperative

Binary regression has many medical applications. Traditionally, the logistic density function is used in binary regression (i.e. logistic regression), and the exposure variable X is treated as a non-stochastic variable. Tiku and Vaughan (1997) showed that the logistic density is restrictive in certain situations and found non- logistic density functions to be more appropriate for some real data examples. Oral and Tiku (2004) extended their approach to the case where the exposure variable X is also a stochastic variable, which is often more realistic from a theoretical as well as a practical point of view. In this study, we further extend this work to multiple stochastic covariates and show that the inclusion of information on the covariates (i.e. the distribution of the covariates) in the likelihood function gives more efficient estimators. Our method is illustrated using data from the Breast Cancer Surveillance Consortium to estimate the effects of stochastic covariates on mammography accuracy.

email: oral.e@ghc.org

COMPARISON OF LOGISTIC REGRESSION VERSUS PROPENSITY SCORING IN BINARY TREATMENT EFFECT ESTIMATION

Yi Huang*, Johns Hopkins University
Karen Bandeen-Roche, Johns Hopkins University
Constantine Frangakis, Johns Hopkins University

Both logistic regression and propensity scoring are used to estimate a treatment effect for observational studies. However, the two approaches may target quite different estimands and formulate quite different questions. The first aim is to clarify the assumptions underlying each modeling approach and thus emphasize the distinction between conditional and marginal treatment effects with respect to model covariates other than the indicator of treatment, made by typical applications of the two approaches. Special attention is put on the non- collapsibility problem associated with different estimands. For example, different applications of the propensity scoring approach should be used when interest shifts from E[Y(1)-Y(0)] to marginal odds ratio. The second aim is to present simulation studies comparing the inferential performance of the two approaches, including propensity scoring methods with different cutting and weighting schemes. Comparisons are carried out for both correctly and incorrectly specified models as well as different sample sizes, case prevalences Pr(Y=1), and strengths of the association between Y and Z. Our work aims to illuminate the practical application of propensity scoring.

email: yhuang@jhsph.edu

A NOTE ON TESTS FOR INTERACTION IN QUANTAL RESPONSE DATA

Melinda M. Holt*, Southeastern Louisiana University
James Stamey, Stephen F. Austin University
John W. Seaman Jr., Baylor University
Dean M. Young, Baylor University

Researchers in a variety of fields have developed tests for interaction in quantal response data using parametric methods, response surface methodology, and large-sample theory. To date, however, few have offered distribution- free methods that could be applied in fixed-dose combination trials. Because the nomenclature for types of joint action is less than standardized across disciplines, deciding which procedure is appropriate for a given study is difficult. Thus a brief synthesis of the terminology and discussion of existing tests of interaction will be provided. Next three new tests to address this issue will be introduced, including the likelihood ratio test, the conditional likelihood ratio test, and a simple new procedure based on the bootstrap. The power of the likelihood ratio test and the new bootstrap test statistic will be examined using an innovative linear extrapolation power-estimation technique due to Boos and Zhang (2000).

email: melinda.holt@selu.edu

AGREEMENT FOR CURVED DATA

Jason Liao*, Merck Research Laboratory

An agreement problem usually involves assessing the concordance of two sets of measurements and the problem covers a broad range of data. In practice, the observations are often curves instead of the traditional points. In this paper, the agreement problem is studied for curved data. Following the rationale in constructing a correlation coefficient curve for heterocorrelaticity, an agreement curve is proposed to measure agreement as a function of the independent variable for curved data. The agreement curve overcomes the drawback when only one index is used in assessing the agreement of two measurements, and it covers all situations including the non-constant mean, non-homogenous variance, and the data range. A real data set is used to demonstrate the approach and to show accurate assessment and information gained if curved data is used.

email: jason_liao@merck.com

77. FROM PROTEOMICS TO CLASSIFICATION TO MEASUREMENT ERROR: CURRENT TOPICS IN PROTEOMICS/GENOMICS

A METHOD FOR SPOT FINDING IN TWO-DIMENSIONAL IMAGES

Jeffrey C. Miecznikowski*, Carnegie Mellon University
William F. Eddy, Carnegie Mellon University
Jonathan S. Minden, Carnegie Mellon University
Kimberly F. Sellers, University of Pennsylvania

Within the study of proteomics, a major goal is to determine the proteins that are present within a tissue sample. Through fluorescence difference gel electrophoresis, scientists have a method for detecting quantitative changes in the amount and type of proteins between two tissue samples. Thus, this technology allows for two samples to be compared and the differences in proteins between the samples to be recorded. In order to automate a process that accurately determines the protein differences between two images we must have an algorithm that locates and quantifies the amount of each protein. Using a median smoother we have developed a method to locate protein spots and assess the amount of protein present in each spot.

email: jcm3@stat.cmu.edu

A CONDITIONAL MOMENT METHOD FOR MODEL SELECTION IN PENALIZED LOGISTIC REGRESSION FOR DISEASE CLASSIFICATION USING MICROARRAY GENE EXPRESSION DATA

J. G. Liao*, University of Medicine and Dentistry, New Jersey

Golub et al (1999) used gene expression to classify/predict between ALL and AML. The use of gene expression for disease classification has been an active area of research since then. The logistic regression provides the natural framework for this as it not only allows for disease classification but also provides the more quantitative probability. Penalized logistic is a popular approach for dealing with the small n and large p problem in this context. The choice of the penalty parameter is the critical and unsolved issue. We propose a conditional moment method. The method is applied to two datasets and compared to existing method for selecting the penalty parameter.

email: jg_liao@yahoo.com

GENE FUNCTION PREDICTION BY A COMBINED ANALYSIS OF GENE EXPRESSION DATA AND PROTEIN-PROTEIN INTERACTION DATA

Guanghua Xiao*, University of Minnesota
Wei Pan, University of Minnesota

Prediction of biological functions of genes is an important issue in basic biology research and has applications in drug discoveries and gene therapies. Previous studies have shown either gene expression data or protein-protein interaction data alone can be used for predicting gene functions. In particular, clustering gene expression profiles has been widely used for gene function prediction. We first propose a new method for gene function prediction using protein-protein interaction data, which will facilitate combining prediction results based on clustering gene expression profiles. We then propose a new method to combine the prediction results based on either source of data by weighting on the evidence provided by each. Using the MIPS gene annotations we show that this new combined analysis provides improved predictive performance over that of using either data source alone.

email: guanghx@biostat.umn.edu

CLASSIX: A NEW CLASSIFICATION METHOD BASED ON A SEPARATION INDEX

Weiliang Qiu*, Channing Laboratory–Brigham and Women's Hospital, Harvard Medical School
Mei-Ling T. Lee, Channing Laboratory–Brigham and Women's Hospital, Harvard Medical School

Linear discriminant analysis (LDA) is one of the commonly used classification methods. It is simple to implement and easy to use. The LDA method is based on an assumption that the covariance matrices of the two classes are the same. In practice, however, data in two classes often have different covariance matrices. In these cases, the performance of LDA may not be satisfactory. To remedy this situation, we propose a new CLASsification method based on a Separation IndeX (CLASSIX for short) to relax the equal-covariance-matrix assumption required by LDA. The decision boundary of CLASSIX is linear. This generalized LDA method has good performance in the analysis of both simulated data sets and a mass spectrometry data set, in comparison with commonly used classification methods, such as LDA, quadratic discriminant analysis (QDA), K-nearest neighbor method (KNN), classification and regression trees (CART), support vector machines (SVM), and random forest (RF).

email: stwxq@channing.harvard.edu

STRUCTURAL EQUATION MODELING OF GENOTYPE BY ENVIRONMENT INTERACTION

Prabhakar Dhungana*, University of Nebraska
Kent M. Eskridge, University of Nebraska
P. S. Baenziger, University of Nebraska
Lenis Nelson, University of Nebraska
W. Stroup, University of Nebraska
Albert Weiss, University of Nebraska
B. T. Campbell, USDA

Coupling structural equation modeling approach with multiplicative interaction components obtained from singular value decomposition of genotypic x environment interaction (GEI), and observed genotypic and environmental covariates can provide a comprehensive way of explaining predictively accurate GEI. The models: SEM- AMMI and SEM-Mixed AMMI developed in this study use pattern-rich GEI considering the first few multiplicative interaction components obtained from singular value decomposition of GEI. We illustrate the uses of these models by analyzing GEI in winter and durum wheat trials. Overall it is concluded that the SEM approach has a distinct advantage over other methods in providing a comprehensive understanding of GEI compensation effects among yield components and in decomposing the total effects of yield components GEI and cross product covariates on grain yield GEI, into their direct and indirect effects. SEM-AMMI allows us to assess the combined effects of genotypic and environmental covariates on yield and yield components GEI when environments and genotypes are fixed whereas SEM-Mixed AMMI allows us to separately model genotypic variability associated with GEI assuming genotypes as random and environments fixed, and thus to identify subsets of genotypic covariates and yield components which are the most sensitive to the environments under study.

email: p_dhungana2002@yahoo.com

MEASUREMENT ERROR MODEL FOR cDNA MICROARRAY AND TIME-TO-EVENT DATA

Jonathan A. L. Gelfond*, University of North Carolina at Chapel Hill
Joseph G. Ibrahim, University of North Carolina at Chapel Hill

It is important to explore and define the relationship between gene expression assays such as microarrays and clinical outcome. Discovery of the association between gene expression and patient outcome rely on statistical models which typically use the expression measurement directly as a predictor. However, a more accurate approach would be to assume that the expression measurement is merely an indication of the underlying true expression level. This measurement error model recognizes the fact that gene expression is not determined with certainty by the microarray, but rather expression level is measured with some error. Experiments involving many tumors will often perform only one microarray assay on each tumor because of cost limitations. The distributions of measurement errors are not identifiable in this case. We propose a model that allows prediction of the magnitude and approximate distribution of the measurement errors in that setting where each tumor is assayed only once. This measurement error model is then included within the framework of a peicewise exponential survival model. Robustness analyses are performed, and the model is applied to a breast cancer study dataset.

email: jgelfond@bios.unc.edu

78. ADVANCED TOPICS IN PROSTATE CANCER MODELING: A MULTIDISCIPLINARY AND INTEGRATED PERSPECTIVE

STATISTICAL MODELING STRATEGIES TO DEFINE THE BIOLOGIC BASIS OF CLINICALLY SIGNIFICANT PROSTATE CANCER

Timothy J. McDonnell*, University of Texas M. D. Anderson Cancer Center

Prostate cancer is among the most common malignancies and a leading cause of cancer mortality among men. Selecting the most appropriate therapy for individual prostate cancer patients is frequently problematic on the basis of currently available predictive information. Further, non- surgical treatments for prostate cancer are seldom curative. We are employing a variety of statistical applications to interrogate datasets obtained from high throughput genome-wide methodologies of clinically annotated human prostate cancer specimens including those obtained from neoadjuvant clinical trials. Candidate gene and validation strategies incorporate the use of tissue microarrays generated from human tissue specimens with clinical annotation. This information is now being incorporated, iteratively, into the design of current clinical trials to accelerate new therapy development.

email: tmcdonne@mdanderson.org

CORRELATING MICROARRAY GENE EXPRESSION MEASUREMENTS WITH GLEASON SCORES AND IDENTIFYING BIOMARKERS TO DISTINGUISH PROSTATE CANCER STAGES

Jing Wang*, University of Texas M. D. Anderson Cancer Center
Kim-Anh Do, University of Texas M. D. Anderson Cancer Center
Sijin Wen, University of Texas M. D. Anderson Cancer Center
Spyros Tsavachidis, University of Texas M. D. Anderson Cancer Center
Timothy J. McDonnell, University of Texas M. D. Anderson Cancer Center
Kevin R. Coombes, University of Texas M. D. Anderson Cancer Center

Identifying tumor subtypes from molecular signatures that can be used as diagnostic and prognostic biomarkers is a significant challenge. Several studies have applied microarrays to prostrate cancer for this purpose. With increased availability of public microarray datasets, it is now possible to combine data across studies. Meta-analysis of microarray data to identify biomarkers is critically important for the development of this technology. Focusing on prostate cancer, we present a novel method for the meta-analysis of microarray data. We develop a useful framework for exploiting bioinformatic and statistical algorithms to overcome some critically important, but difficult, problems for biological and clinical applications of the technology. Based on our analysis of data from two institutions using different platforms, we constructed a probabilistic model that uses 7 genes to distinguish tumors with good prognostic Gleason score (6 or lower; N=38) from tumors with poor prognostic Gleason score (8 or higher; N=13), with 100% accuracy on the training set. When applied to 51 tumors from patients with intermediate Gleason score, the model separated them into good or poor prognostic categories that were associated with Gleason grade 3+4 or 4+3, respectively (Fisher exact test; p = 0.033). Biological confirmation using other technologies (RT-PCR, IHC) is continuing.

email: jingwang@mdandson.org

COMBINING LONGITUDINAL STUDIES OF PSA

Lurdes YT Inoue*, University of Washington
Ruth Etzioni, Fred Hutchinson Cancer Research Center
Elizabeth Slate, Medical University of South Carolina
Christopher Morrell, Loyola College in Maryland
David F. Penson, Keck School of Medicine and University of Southern California/Norris Cancer Center

Prostate-specific antigen (PSA) is a biomarker commonly used to screen for prostate cancer. Several studies have examined PSA growth rates prior to prostate cancer diagnosis. However, the resulting estimates are highly variable. In this talk we present a non-linear Bayesian hierarchical model to combine longitudinal data on PSA growth from three different studies. Our model enables novel investigations into patterns of PSA growth that were previously impossible due to sample size limitations. The goals of our analysis are two-fold. First, to characterize growth rates of PSA accounting for study differences. Second, to investigate the impact of clinical covariates such as advanced disease and unfavorable histology on PSA growth rates.

email: linoue@u.washington.edu

BAYESIAN NETWORKS FOR PROSTATE CANCER MODELING

Bradley M. Broom*, University of Texas M. D. Anderson Cancer Center
Devika Subramanian, Rice University

Two key problems in prostate cancer are discovering genetic regulatory networks that characterize a state of the disease and understanding their evolution with disease progression. Specifically, we want to identify the differences between the regulatory networks for low and high grade prostate cancers, and between primary and metastatic disease. Several hundred genes are known to be differentially expressed in various stages of prostate cancer. Understanding of the system level interactions between all of these genes and their role in disease onset and progression is incomplete. Bayesian networks are useful for representing and learning about system wide gene regulation. Their ability to posit hidden variables is important for inferring gene regulatory networks. The challenges for constructing Bayesian models of gene regulation include the large number of genes involved, the need to consider a number of networks super-exponential in the number of genes, and the limited, noisy data that makes reliable discrimination between alternative networks difficult. In this paper we describe our approach to addressing these challenges in the context of prostate cancer gene expression data.

email: broom@odin.mdacc.tmc.edu

---

79. STATISTICAL METHODS IN QUANTITATIVE GENETICS AND GENOMICS

BAYES ESTIMATORS FOR MOLECULAR GENEALOGY

Bruce Walsh*, University of Arizona

Genealogists wishing to forge connections between potentially related individuals in the absence of any paper document trait are turning to DNA. While forensic studies use a set of roughly a dozen unlinked autosomal markers, these are inappropriate for detecting relatives separated by more than one or two generations, as recombination quickly removes any signal. In contrast, completely linked markers (such as on the non-recombining arm of the Y chromosome) have their relatedness information decay on order of the mutation rate (1/500 vs. 1/2 for recombination). Estimates of the time to the most recent common ancestor (TMRCA) for two individuals based on Y marker information are developed. Likelihood estimators are developed, but shown to have problems, especially given the small sample size (number of markers). Population genetics provides a natural prior for TMRCA (geometric with parameter 1/ effective population size) and Bayesian estimators are developed using this prior. Explicit posteriors are computed under both the infinite alleles model and the stepwise mutational model.

email: jbwalsh@u.arizona.edu

LISTS OF LISTS: A HIERARCHICAL INFERENCE PROBLEM FROM GENOMICS

Michael A. Newton*, University of Wisconsin–Madison

Consider a microarray study comparing gene expression between two cellular conditions. Having scored genes for differential expression (DE), the investigator wishes to identify sets of genes that are enriched for DE genes. Gene Ontology (GO) annotations provide a case in point; each set is a collection of genes that are associated with a common biological process, molecular function, or cellular localization. The hypergeometric distribution has been used to measure enrichment of DE genes in a GO annotation. One considers the cross classification of genes according to whether or not they are on the DE list and whether or not they have the annotation. A problem with this `selection' approach is that it reduces quantitative information about DE to indicators of whether or not genes are on the DE list. I will present an alternative `averaging' approach that uses additional information from the DE analysis. Neither approach is uniformly superior. Selection is preferred when the DE effect is large and the DE extent (proportion of affected genes in the annotation) is small. Averaging is preferred when the effect is small but the extent is large. In support of these claims I will present evidence from theoretical and empirical calculations.

email: newton@stat.wisc.edu

CLASS DISCOVERY AND CLASSIFICATION OF TUMOR SAMPLES USING MIXTURE MODELING OF GENE EXPRESSION DATA—A UNIFIED APPROACH

Shili Lin*, Ohio State University

DNA microarray technology has been increasingly used in cancer research. In the literature, discovery of putative classes and classification to known classes based on gene expression data have been largely treated as separate problems. This article offers a unified approach to class discovery and classification, which we believe is more appropriate, and has greater applicability, in practical situations. We model the gene expression profile of a tumor sample as from a finite mixture distribution, with each component characterizing the gene expression levels in a class. The proposed method was applied to a leukemia dataset, and good results are obtained. With appropriate choices of genes and preprocessing method, the number of leukemia types and subtypes is correctly inferred, and all the tumor samples are correctly classified into their respective type/subtype. Further evaluation of the method was carried out on other variants of the leukemia data and a colon dataset. The program implementing the method can be downloaded freely at http://www.stat.ohio-state.edu/~statgen/SOFTWARE/DNC-MIX/. This is joint work with Roxana Alexandridis and Mark Irwin.

email: shili@stat.ohio-state.edu

80. STATISTICAL METHODS FOR ANALYSIS OF GENE-ENVIRONMENT INTERACTION

SEMIPARAMETRIC METHODS FOR ESTIMATING GENE-ENVIRONMENT INTERACTION PARAMETERS FROM CASE-CONTROL STUDIES

Christie Spinka*, University of Missouri–Columbia
Raymond Carroll, Texas A&M University
Nilanjan Chatterjee, National Cancer Institute

Case-control studies of unrelated subjects are now widely being used to study the role of genetic susceptibility and gene-environment interactions in the etiology of complex diseases. Exploiting an assumption of gene-environment independence, and treating the distribution of the environmental exposures to be completely non-parametric, we have recently developed an efficient retrospective maximum-likelihood method for analysis of case-control studies (Chatterjee and Carroll, Biometrika, in press). In this article, we develop an extension of the retrospective maximum-likelihood approach to studies where genetic effects are modeled in terms of "haplotypes", the combination of alleles at multiple loci in a single chromosome, but the exact haplotype configuration in two chromosomes of some subjects cannot be derived with certainty from available locus-specific genotype data. We use a profile-likelihood technique and an appropriate EM algorithm to derive a relatively simple procedure for parameter estimation. We also describe an alternative estimating equation-based approach that is less sensitive to the gene-environment independence assumption. We discuss how the latter approach contrasts with some of the recently proposed "prospective'' methods that could be inconsistent even when the underlying gene-environment independence and Hardy-Weinberg-Equilibrium assumptions are valid for the underlying population. The methods are illustrated through simulation studies and real data examples.

email: chattern@mail.nih.gov

A METHOD FOR USING NUCLEAR FAMILIES TO IDENTIFY GENE BY ENVIRONMENT INTERACTION

Emily O. Kistner*, National Institute of Environmental Health Sciences
Clarice R. Weinberg, National Institute of Environmental Health Sciences
Claire Infante-Rivard, McGill University

For complex human diseases, both exposures and genetic polymorphisms can influence susceptibility. We consider a design where cases and their parents are genotyped and demonstrate that an approach similar to one suggested for testing linkage and association between genetic markers and quantitative traits (Kistner and Weinberg 2004) can be used to test gene by environment interactions. We presume a multiplicative null. Thus, we ask whether the relative risk (s) associated with the exposure vary across the inherited genotypes. The same polytomous logistic framework we proposed to identify genes related to a quantitative trait applies, except that the environmental exposure is substituted for the quantitative trait. By conditioning on parental genotypes, both approaches allow for genetic population admixture and nonmendelian transmission between parents and their offspring. The proposed method also allows multiple cases per family by using a weighted score function instead of a likelihood ratio test. In addition, missing parental genotype data are incorporated through a multiple imputation procedure. The proposed gene- environment interaction test will be demonstrated by applying it to a sample of newborns with intrauterine growth restriction. The test will consider interactions between maternal smoking and genes that metabolize smoking by-products.

email: kistner@niehs.nih.gov

HAPLOTYPES IN STUDIES OF GENE BY ENVIRONMENT INTERACTION

Peter Kraft*, Harvard University

I review several analytic strategies using haplotypes inferred from multilocus genotypes in the context of case-control studies of gene x environment interaction. These strategies can be divided into two categories: single imputation and marginal-over-the-missing-data approaches. Simulation and analytic results suggest that both approaches can provide accurate estimates of haplotype odds ratios in regions of high linkage disequilibrium and limited haplotype diversity. However, outside of such regions, both approaches can show noticeable bias. I discuss the implications of this observation for the design of 'haplotype-tagging' studies of candidate genes. In particular, I argue 'haplotype-tagging' analyses should be restricted to regions that show evidence of high linkage disequilibrium and limited haplotype diversity.

email: pkraft@hsph.harvard.edu

81. ISOTONIC METHODS IN TOXICOLOGY & RISK

BAYESIAN METHODS FOR ASSESSING ORDERING IN HAZARD FUNCTIONS

Laura H. Gunn*, Georgia Southern University
David B. Dunson, National Institute of Environmental Health Sciences

In biomedical studies that collect event time data, it is often appropriate to assume non-decreasing hazards across dose groups, though dose effects may vary with time. Motivated by this application, we propose a Bayesian approach for order restricted inference using a non- proportional hazards model with time-varying coefficients. In order to make inferences on equalities versus increases in hazard functions, a prior is chosen for the time-varying coefficients that assigns positive probability to no dose effect while restricting coefficients to be non-negative. By using a high dimensional piecewise constant model and smoothing the functions by coupling Markov beta and gamma processes, we obtain a flexible and computationally tractable approach for identifying sets of dose and age values at which hazards are increased. This approach can also be used to estimate dose response and survival curves. The methods are illustrated through application to data from a toxicology study.

email: lgunn@georgiasouthern.edu

USE OF HISTORICAL CONTROLS IN SURVIVAL-ADJUSTED QUANTAL RESPONSE TESTS FOR COMPARING TUMOR INCIDENCE RATES

Shyamal D. Peddada*, National Institute of Environmental Health Sciences
Gregg E. Dinse, National Institute of Environmental Health Sciences
Grace E. Kissling, National Institute of Environmental Health Sciences

In animal carcinogenicity studies, such as those conducted by the National Toxicology Program (NTP), researchers are often interested in comparing tumor rates across dose- groups to detect dose-related trends. Over the years the NTP has compiled a large database of various tumors for control animals. Researchers are interested in using this historical control database in analyzing the current data, in particular, to improve the power of a test procedure when dealing with rare tumors. Difficulties are often encountered when a small number of tumors are seen in the high dose groups and no tumors are seen in the concurrent control group. The number of tumors observed in the high dose groups may be small enough that standard procedures fail to detect a significant trend, yet in the historical control database the tumor is very rare. At the moment there is no satisfactory methodology that incorporates historical control data and survival times when detecting dose-related trends in the current experimental data. The focus of this talk is to present a new methodology for detecting dose-related trends in tumor rates that makes use of the historical control database. The proposed methodology is based on the recent trend test of Peddada, Dinse, and Haseman (J.R.S.S. - Ser. C, 2005). We discuss the performance of the proposed procedure and illustrate it using an NTP example.

email: peddada@niehs.nih.gov

USING ISOTONIC REGRESSION TO IDENTFY THE IDEAL RECALL RATE IN SCREENING MAMMOGRAPHY

Michael J. Schell*, University of North Carolina at Chapel Hill
Bahjat F. Qaqish, University of North Carolina at Chapel Hill
Bonnie C. Yankaskas, University of North Carolina at Chapel Hill
Monique A. Amamoo, Shaw University
William E. Barlow, University of Washington

Isotonic regression is an underutilized statistical procedure. In many regression settings, retaining monotonicity, while relaxing the assumption of linearity is desirable in regression modeling. In this paper, the sensitivity of screening mammography is estimated as an isotonic function of the recall rate (the fraction of women recalled for additional testing to ascertain the presence of breast cancer). The reduced monotonic regression procedure of Schell and Singh (JASA, 1997) is stopped early so that the number of isotonic level sets is only somewhat reduced. Then, piecewise linear fitting is applied to the weighted centers of the new level sets. From this non- decreasing piecewise linear fit, the minimum number of 'extra work-ups per extra cancer found' (EWUPECF) can be calculated for various recall rates. The 'ideal' recall rate to target for in screening mammography practice, then, depends upon the EWUPECF values. This applied study uses data from the Breast Cancer Surveillance Consortium.

email: mjschell@earthlink.net

82. NOVEL ENVIRONMENTAL APPLICATIONS OF SPATIAL STATISTICS

SINGLE- AND MULTI-RESOLUTION COREGIONALIZED MODELS FOR SPATIALLY-VARYING GROWTH CURVES

Sudipto Banerjee* and Gregg A. Johnson, University of Minnesota

Weed growth in agricultural fields constitutes a major deterrent to the growth of crops, often resulting in low productivity and huge losses for the farmers. Therefore, proper understanding of patterns in weed growth is vital to agricultural research. Recent advances in Geographical Information Systems (GIS) now allow geocoding of agricultural data, which enables more sophisticated spatial analysis. Our current application concerns the development of statistical models for conducting spatial analysis of growth patterns in weeds. Our data comes from an experiment conducted in Waseca, Minnesota, that recorded growth of the weed Setaria spp. We capture the spatial variation in Setaria spp. growth using spatially-varying growth curves. An added challenge is that these designs are spatially replicated, with each plot being a lattice of sub-plots. Therefore, spatial variation may exist at different resolutions - a macro level variation between the plots and micro level variation between the sub-plots nested within each plot. We develop a Bayesian hierarchical framework for this setting. Flexible classes of models result which are fitted using simulation-based methods.

email: sudiptob@biostat.umn.edu

SPATIAL ANALYSIS OF SEA TURTLE NESTING AT JUNO BEACH, FLORIDA

Lance A. Waller*, Emory University
Traci Leong, Emory University
Andrew Barclay, Emory University
Bud Howard, Department of Environmental Resources Management, Palm Beach County

We present a spatial statistical analysis of nesting patterns of loggerhead and green turtles along Juno Beach, Florida for the years 1997-2000. Data include approximately 8,000-10,000 emergence sites per year, located with sub-meter accuracy via global positioning system (GPS) technology. We treat nesting patterns as spatial point processes and compare estimates of the first- and second-order features of the observed patterns to address questions regarding the impact of a newly-constructed fishing pier on nesting patterns, the impact of beach renourishment on nesting patterns, and inter-species differences in nesting behavior.

email: lwaller@sph.emory.edu

MODELING MULTIVARIATE SPATIAL VARIABLES

Hao Zhang*, Washington State University

In many environmental, agricultural and epidemiological studies, multiple spatial variables are observed. Not only each of these variables is spatially correlated at different locations, but also these variables may be cross- correlated. The correlation structure of these variables can be described through the multivariate covariogram or variogram, which is employed in cokriging for example and helps describe how these variables relate to each other. A few models for multivariate covariogram have been proposed. However, estimation of the parameters in the multivariate covariogram models is understudied. We show how the maximum likelihood estimators (MLEs) can be calculated through the EM algorithm for an important class of multivariate covariogram. Furthermore, the MLEs automatically satisfy the constraints necessary for multivariate covariogram. We also present results of analyses on some real datasets.

email: zhanghao@wsu.edu

HIERARCHICAL BAYESIAN MODELING OF INVASIVE SPECIES

Christopher K. Wikle*, University of Missouri–Columbia
Mevin B. Hooten, University of Missouri–Columbia

Ecological processes often exhibit very complicated spatiotemporal dependencies. To understand and eventually predict such complicated processes, we must make use of available scientific insight, data, and theory, in a modeling framework that honestly accounts for uncertainties in each. Specifically, we are concerned with modeling invasive species in this context. In general, a successful invasion includes four stages: introduction, establishment, range expansion and saturation. We consider these stages as part of a hierarchical Bayesian framework, in which discrete space-time dynamics drive the invasion. An invasive species data set will be used to demonstrate the model and methodology.

email: wiklec@missouri.edu

83. NONPARAMETRICS

COMPARISON OF CURVES BASED ON A CRAMER-VON-MISES STATISTIC

Hua Liang*, St. Jude Children's Research Hospital

We propose a test to compare two curves and investigate the limiting behavior of the test statistic. The test can detect the local alternative converging to the null at the parametric rate. To calculate the critical values, the bootstrap resample technique is employed. To show how the test works, we conduct a simulation experiment to study the level and power of the bootstrap test. We also compare the test with those proposed by Hall and Hart (1990) and Young and Bowman (1995) in the simulation study. The tests are further used to analyze a real data set.

email: hua.liang@stjude.org

CANONICAL CORRELATES FOR FOUR SETS OF FUNCTIONAL DATA CURVES

Peter M. Meyer*, Rush University Medical Center
Sue Leurgans, Rush University Medical Center

Leurgans et al. (1993) described the use of canonical correlation in the setting of functional data analysis. Kettenring (1971) discussed various approaches for extending canonical correlation to several sets of variables. We compare the various approaches for finding canonical correlates in the functional data analysis setting for four sets of hormone curves, and present implications for other collections of three or more sets of variables. Data are from the Daily Hormone Study of Studies in Women's Health Across the Nation (SWAN). S. Leurgans, R. Moyeed, and B. W. Silverman. Canonical correlation analysis when the data are curves. Journal of the Royal Statistical Society, B, 55(3):725-740, 1993. J. R. Kettenring. Canonical analysis of several sets of variables. Biometrika, 58(3):433-451, 1971.

email: pmeyer@rush.edu

ON SOME TESTS OF THE COVARIANCE MATRIX UNDER GENERAL CONDITIONS

Arjun K. Gupta, Bowling Green State University
Jin Xu*, University of California, Riverside

We consider the problem of testing the hypothesis about the covariance matrix of random vectors under the assumptions that the underlying distributions are nonnormal and the sample size is moderate. The asymptotic expansions of the null distributions are obtained up to n^(-1/2). It is found that in most cases the null statistics are distributed as a mixture of independent chi-square random variables with degree of freedom one (up to n^(-1/2)) and the coefficients of the mixtures are functions of the fourth cumulants of the original random variables. We also provide a general method to approximate such distributions based on a normalization transformation.

email: jin.xu@ucr.edu

NONPARAMETRIC ESTIMATION OF STABLE EXPONENT

Zhaozhi Fan*, University of New Hampshire

Assume that $X_{1},X_{2},\cdots,X_{n}$ is a sequence of i.i.d. random variables with α-stable distribution ($\alpha \in (0,2]$, the stable exponent, is the unknown parameter). We construct nonparametric estimators for $ \alpha$ by minimizing the Kolmogorov distance or the Cramer-von Mises distance between the empirical distribution function G_{n}, and a class of distributions defined based on the sum-preserving property of stable random variables. These estimators can also be obtained by minimizing a U-statistic estimate of an empirical distribution function involving the stable exponent. They share the same invariance property with the maximum likelihood estimates. In this paper we prove the strong consistency of the estimators. We prove a large deviation principle. Simulation study shows that the new estimators are competitive to the existing ones and perform very closely even to the maximum likelihood estimator.

email: zfan@unh.edu

NONPARAMETRIC DECONVOLUTION, TRADITIONAL AND NONTRADITIONAL POOLED DESIGN

Albert Vexler*, National Institute of Child Health and Human Development, National Institutes of Health
Aiyi Liu, National Institute of Child Health and Human Development, National Institutes of Health
Enrique F. Schisterman, National Institute of Child Health and Human Development, National Institutes of Health

Let a sequence of independent and identically distributed random variables X with an unknown density f be given. An often-occurring problem in statistics is that we have observations Xp, which are equal to the partial sums of variables of interest Xs. The objective of this research is to propose and examine a methodology developed to estimate f nonparametrically based on samples of observations of Xp. An important task of such studies is nonparametric ROC curve analysis for biomarkers based on pooled assessments

email: vexlera@mail.nih.gov

POINTWISE COMPARISONS FOR FUNCTIONAL DATA INFERENCE

Dennis Cox, Rice University
Jong Soo Lee*, Rice University

We will consider the problem of hypothesis testing for functional data. Although there exist methods of hypothesis testing in the functional data setting, most of them are concerned with global testing (similar to an F-test in ANOVA). We propose a pointwise testing procedure which can be used either by itself or in conjunction with global tests. First, we briefly survey various multiple comparison methods applied to pointwise tests. Then we arrive at a randomization-based method of Westfall and Young, which is shown to work well and has desirable theoretical properties. The talk will conclude with discussion of implementation issues and possible future applications and improvements.

email: jslee@stat.rice.edu

SIMULTANEOUS ESTIMATION OF INDIVIDUAL PATIENTS' RESPONSES

Jin Zhu*, Novartis Pharmaceuticals

In clinical trials it is very often that effects in subsets need to be estimated, for example, effects in each center or exacerbation rate of each patient. In order to estimate the effects accurately, a reasonable sample size is required from each subset or patient. However this may not always be feasible, for example, in the latter example above. Compound decision theory, a robust form of empirical Bayes decision theory, can be used in this type of situation. In this presentation, compound decision theory will be briefly described with emphasis on the linear compound decision theory. Simulation and application on real data will be illustrated.

email: jin.zhu@pharma.novartis.com

84. STATISTICAL METHODS IN SCREENING

STATISTICAL EVALUATION OF INTERNAL AND EXTERNAL MASS CALIBRATION LAWS UTILIZED IN FOURIER TRANSFORM ION CYCLOTRON RESONANCE MASS SPECTROMETRY

Ann L. Oberg*, Mayo Clinic and Foundation
David C. Muddiman, Mayo Clinic and Foundation
Terry M. Therneau, Mayo Clinic and Foundation
Jeanette E. Eckel-Passow, Mayo Clinic and Foundation

Many scientists in cancer research are searching for biomarkers for use in early screening for cancer in hopes of reducing cancer mortality. Mass spectrometry-based proteomic methods such as Surface Enhanced Laser Desorption Ionization (SELDI) and Matrix Assisted Laser Desorption Ionization (MALDI) coupled with time-of-flight mass analyzer technology have the ability to interrogate an individual's proteome. An alternative biomarker discovery platform is based on the coupling of liquid chromatography (LC) with Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) using an electrospray ionization (ESI) interface. FT-ICR-MS affords wide dynamic range, high mass measurement accuracy (MMA), unparalleled mass resolving power and high sensitivity. The success of any biomarker discovery platform is inherently dependent on the extent of variability including biological, disease heterogeneity, and, although not yet well quantified in proteomics, sample handling. In an effort to maximize the high MMA that FT-ICR-MS affords, we evaluated several mass calibration laws and assessed the impact of various factors on MMA in a carefully designed experiment. Due to potential interference of internal calibrant ions with analyte species, both internal and external calibration laws were investigated. Results and potential implications will be presented in the context of using LC-FT-ICR-MS as a biomarker discovery platform.

email: oberg.ann@mayo.edu

ON CRITERIA FOR EVALUATING MODELS OF ABSOLUTE RISK

Ruth M. Pfeiffer*, National Cancer Institute, NIH
Mitchell H. Gail, National Cancer Institute, NIH

Absolute risk is the probability that an individual who is free of a given disease at an initial age, a, will develop that disease in the subsequent interval (a, t]. Absolute risk is reduced by mortality from competing risks. Models of absolute risk that depend on covariates have been used to design interventions studies, to counsel patients regarding their risks of disease, and to inform clinical decisions, e.g. whether or not to take chemoprevention for breast cancer. Several general criteria have been used to evaluate absolute risk models, including how well the model predicts the observed numbers of events in subsets of the population ("calibration"), and "discriminatory power", measured by the concordance statistic. We develop specific loss function-based criteria for two applications, namely whether to screen a population to select subjects for further evaluation or treatment and whether to use a preventive intervention that has both beneficial and adverse effects. We find that high discriminatory power is much more crucial in the screening application than in the preventive intervention application. These examples indicate that the usefulness of a general criterion such as concordance depends on the application, and that using specific loss functions can lead to more appropriate assessments.

email: pfeiffer@mail.nih.gov

ASSESSING RELATIVE ACCURACY OF SCREENING TESTS IN A RANDOMIZED PAIRED SCREEN POSITIVE DESIGN

Todd A. Alonzo*, University of Southern California
John M. Kittelson, University of Colorado Health Sciences Center

For ethical reasons, studies evaluating the accuracy of two binary screening tests for cervical cancer often employ a paired screen positive (PSP) design where only women who screen positive on at least one of the two screening tests receive the gold standard (biopsy). In settings where a PSP design is not technically feasible, we propose a randomized paired screen positive (RPSP) design where women are randomized to receive one of the two screening tests and only receive the other screening test and gold standard if the first screening test is positive. Maximum likelihood estimators of the relative accuracy of the two screening tests and corresponding confidence intervals are developed for RPSP studies. Simulation studies are used to assess the small sample bias of the point estimators and to assess the coverage probabilities of the confidence intervals. Simulations are also used to compare the efficiency of a RPSP design to a non-randomized PSP design and to an unpaired screen positive design.

email: talonzo@childrensoncologygroup.org

MODELING THE RELATIONSHIP BETWEEN SENSITIVITY AND SOJOURN TIME IN PERIODIC CANCER SCREENING

Dongfeng Wu*, Mississippi State University
Gary L. Rosner, University of Texas M. D. Anderson Cancer Center
Lyle D. Broemeling, University of Texas M. D. Anderson Cancer Center

This research extends previous probability models for periodic breast cancer screening examinations. The specific aim is to study the relationship between screening sensitivity and sojourn time distribution in the setting of a periodic screening scenario. Literature review shows that they may be negative correlated. We apply our method to the HIP study of female breast cancer and provide statistical inference on the issue.

email: dw183@ra.msstate.edu

ESTIMATION OF SENSITIVITY AND SOJOURN TIME IN BREAST CANCER SCREENING STUDIES

Xiuyu J. Cong*, Rice University
Yu Shen, University of Texas M. D. Anderson Cancer Center

This study investigates statistical approaches to quantitatively describing the effect of age and tumor size at detection on screening sensitivity and sojourn time distribution in breast cancer screening studies. Such an investigation is directly motivated by the need to understand the inherent relationships between the disease natural history and screening programs. We incorporate the age and size effect through generalized linear models under a progressive disease modeling framework and obtain the corresponding parameter estimates using the maximum likelihood method. Extensive simulation studies show that the estimators have reasonable accuracy. The proposed methods are illustrated using data from two large breast cancer screening trials. The results show that the screening sensitivity increases with age and tumor size at screening exams based on these two trials.

email: xcong@stat.rice.edu

A LINEAR MIXED MODEL WITH HETEROSCEDASTIC COVARIATE MEASUREMENT ERROR

Liang Li*, Cleveland Clinic Foundation
Tom Greene, Cleveland Clinic Foundation

Glomerular filtration rate (GFR) is an important clinical measurement in studies of chronic renal diseases. It is found that GFR is measured with error and the magnitude of the error is dependent on the value of the GFR. In a longitudinal study of follow-up clinical outcomes using baseline GFR measurements, failure to account for the measurement error will result in bias in regression coefficients. We use a classical additive error model where the error variance depends on the unobserved true GFR. The parameters are estimated by a set of unbiased estimating equations. Simulations and an example will be given to illustrate the method.

email: lli@bio.ri.ccf.org

LATENT CLASS REGRESSION ON LATENT FACTORS

Jia Guo*, University of Minnesota
Melanie M. Wall University of Minnesota
Yasuo Amemiya, IBM T. J. Watson Research Center

In the research of public health, psychology, and social sciences, many research questions investigate the relationship between a categorical outcome variable and continuous predictor variables. The focus of this paper is to develop a model to build this relationship when both the categorical outcome and the predictor variables are latent (i.e. not observable directly). This model extends the latent class regression model so that it can include regression on latent predictors. Maximum likelihood estimation is used and two numerical methods for performing it are described: the Monte Carlo Expectation and Maximization algorithm (MCEM) and Gaussian quadrature followed by quasi-Newton algorithm. A simulation study is carried out to examine the behavior of the model under different scenarios. A data example involving adolescent health is used for demonstration where the latent classes of eating disorders risk are predicted by the latent factor body satisfaction.

email: jiaguo@biostat.umn.edu

MODEL ROBUSTNESS IN STRUCTURAL MEASUREMENT ERROR MODELING

Xianzheng Huang*, North Carolina State University
Leonard A. Stefanski, North Carolina State University
Marie Davidian, North Carolina State University

Models involving unobservable latent quantities, such as structural measurement error models and so- called 'joint' models for longitudinal and time-to-event data, are widely used in a host of applications. Provided that the model for the latent variable is correctly specified, likelihood-based approaches are appealing because they lead to consistent and effcient inference. However, intuition suggests that misspecification of this model may compromise such inference, although some recent empirical studies have exhibited striking robustness to the assumption on the latent variable. The data analyst faces the diffculty that the extent to which inference may be sensitive to the choice of model for unobservable latent variables is not known in a given problem. Techniques for studying and diagnosing robustness in these models would thus be invaluable. We present a framework for assessing model robustness in the class of structural latent variable models, focusing on the particular subclass of structural measurement error models, and propose practical strategies for diagnosing misspecification of the model for the true predictor, the latent variable for this subclass. The methods are illustrated via several analytic examples and by application to simulated data and a study of coronary heart disease.

email: xhuang@stat.ncsu.edu

ON CORRECTED SCORE APPROACH FOR PROPORTIONAL HAZARDS MODEL WITH COVARIATE MEASUREMENT ERROR

Xiao Song*, University of Washington
Yijian Huang, Emory University

In the presence of covariate measurement error with the proportional hazards model, several functional modeling methods have been proposed. These include the conditional score estimator (Tsiatis and Davidian, 2001), the parametric correction estimator (Nakamura, 1992) and the nonparametric correction estimator (Huang and Wang, 2000, 2003) in the order of weaker assumptions on the error. Although they are all consistent, each suffers from potential difficulties with small samples and substantial measurement error. In this article, upon noting that the conditional score and parametric correction estimators are asymptotically equivalent in the case of normal error, we investigate their relative finite sample performance and discover that the former is superior, which may be explained by the unbiasedness of its estimating equation. This finding motivates a general refinement approach to parametric and nonparametric correction methods. The refined correction estimators are asymptotically equivalent to their standard counterparts, but have improved numerical properties. Simulation results and application to an HIV clinical trial are presented.

email: songx@u.washington.edu

A BAYESIAN ADJUSTMENT OF COVARIATE MISCLASSIFICATION WITH CORRELATED BINARY OUTCOME DATA

Dianxu Ren*, University of Pittsburgh
Roslyn A. Stone, University of Pittsburgh

Estimates of association between an outcome variable and misclassified covariates tend to be biased when the usual methods of estimation that ignore the classification error are applied. Available methods to account for misclassification often require the use of a gold standard (i.e, a validation subsample). But in practice, a gold standard may be unavailable or impractical. We propose a Bayesian approach to adjust for misclassification in a binary covariate in logistic and random effect logistic models when a gold standard is not available. This Markov Chain Monte Carto (MCMC) approach uses two imperfect measures of a dichotomous exposure under the assumption of conditional independence and non-differential misclassification. We illustrate the proposed approach to adjust for misclassification with respect to oxygenation status in a multi-center trial of patients with pneumonia. We validate the approach with a simulation study. Ignoring misclassification produces downwardly biased estimates and underestimate uncertainty

email: dir8@pitt.edu

EVALUATING AND CORRECTING GUESS EFFECT IN IMPERFECT DOUBLE-BLINDED CLINICAL TRIALS

Jianfeng Cheng*, Columbia University
Eva Petkova, Columbia University

Failure to maintain blindness is a common phenomenon in double-blinded clinical trials testing the efficacy of psychiatric medications, which may affect the magnitude of estimated drug effects. We propose a method to assess the presence of possible rater's bias due to treatment guess in psychiatric clinical trials. We also propose the estimators for both guess effect and treatment efficacy by correcting for such guess effect. The performance of the estimators is illustrated by simulation studies. The proposed method is applied to real data from antidepressant clinical trials. Two conclusions can be drawn based on the results from the applications: (1) If guess plays a critical role in clinician's rating, it could create the appearance of treatment effect where such effect does not exist; (2) If guess plays a moderate role in clinician's rating, it could magnify the estimated treatment effect. The proposed method allows one to re-evaluate previously conducted clinical trials, providing an applicable tool to measure and correct the potential bias.

email: jc794@columbia.edu

A LATENT VARIABLE MODEL FOR MEASUREMENT ERROR CORRECTION USING REPLICATE DATA

Sohee Park*, Harvard University
Louise M. Ryan, Harvard University
John Meeker, Harvard University
Russ Hauser, Harvard University

It is well known that when the exposure variable is measured with error, the estimated relationship between exposure and outcome can be seriously biased unless appropriate adjustments are made. To assess within-person variability in the exposure, replicate measurements are often obtained. Using an estimating equations approach, we generalize the standard methods for replicate data to allow for various covariance structures among replicates, e.g., where exposure variable has short-term and long-term variation so that the repeated measures are not simply replicates of each other and replicates are observed on only a subset of study subjects. By simulation studies, we compare the performance of correctly estimating the variance of the parameter for the association between exposure and outcome: 1) using a naive variance estimator from a two-stage fitting method; 2) using Bootstrap samples; and 3) using an estimating equations approach. We also apply these methods to real data from a study of pesticide exposure and male reproductive health outcome.

email: shpark@hsph.harvard.edu

---

86. FRAILTY MODELS

A SHARED GAMMA FRAILTY MODEL FOR DEPENDENT FAILURE AND TRUNCATION TIMES

Emily C. Martin*, Harvard University
Rebecca A. Betensky, Harvard University

Randomly truncated survival data arise when the failure time is observed only if it falls within a subject-specific truncation interval, usually delimited by a single truncation time. Censored and truncated survival data arise when the failure time is not observed exactly, but known to fall within a subset of the truncation interval. Most estimators of the survival function based on such data rely on the key assumption of quasi-independence of failure and truncation times, as well as the usual assumption of independent censoring. Methods are needed for analyzing truncated survival data when a test of quasi-independence fails. We propose a semiparametric shared frailty model that assumes the failure and truncation times are quasi-independent, conditional on a gamma distributed frailty. Estimates of the baseline hazard functions, covariate and frailty effects, and the association parameter are obtained via a modified EM algorithm. For left-truncated survival data, the approach yields Nelson-Aalen-like estimates of the survival function and the cumulative distribution function for truncation time, adjusted for dependent truncation. We demonstrate the method on data sets with both positive and negative dependence and assess performance via simulation.

email: emartin@hsph.harvard.edu

PARAMETRIC FRAILTY MODELS FOR QUALITY OF LIFE IN ONCOLOGY

Andrea B. Troxel*, University of Pennsylvania School of Medicine

Oncology studies often collect information on both clinical and quality of life (QOL) events. The QOL events are defined as occurring when a repeatedly measured scale or diary item surpasses a threshold of interest. Multivariate survival methods are an appealing analysis tool, but must be modified to handle unique features of QOL data. These include grouping of QOL event data, and more importantly, asymmetric dependent censoring induced on the QOL event by the survival event. We propose parametric survival models incorporating a shared gamma frailty parameter linking the two event types. The likelihood incorporates the asymmetry inherent in the event structure. This both accomodates the dependence and allows for estimation of the correlation between event types, which is of great interest in oncology. We describe the estimation process and demonstrate with an example.

email: atroxel@cceb.upenn.edu

LIKELIHOOD RATIO TEST FOR THE VARIANCE COMPONENT IN A SEMI-PARAMETRIC SHARED GAMMA FRAILTY MODEL

Xin Zhi*, Eli Lilly and Company
Lynn Eberly, University of Minnesota
Patricia Grambsch, University of Minnesota

To accommodate the intracluster correlation among survival times in clustered time-to-event data, a cluster random effect (frailty) can be incorporated into the Cox proportional hazards model. We consider the problem of testing whether the frailty variance is 0 in a semi- parametric shared gamma frailty model using likelihood ratio test (LRT). With a nonnegativity restriction on the variance parameter, the LRT is one-sided and with null value on the boundary of parameter space. Under a semi- parametric setting, the asymptotic distribution of this test has not been developed. In this paper, we show that the null distribution of this one-sided LRT statistic converges to a 50:50 mixture of χ^2_0 and χ^2_1 as the number of clusters goes to infinity. Simulation results show that when the number of clusters is large, the null distribution of the LRT statistic is very well approximated by the asymptotic chi-square mixture; when the number of clusters is small, this approximation is off and the LRT based on the asymptotic distribution tends to be conservative in favor of the null. The power of the LRT varies according to the number of clusters and subjects per cluster, as well as the magnitude of the frailty variance.

email: zhi_xin@lilly.com

BAYESIAN SEMIPARAMETRIC DYNAMIC FRAILTY MODELS FOR MULTIPLE EVENT TIME DATA

Michael L. Pennell*, University of North Carolina at Chapel Hill and National Institute of Environmental Health Sciences
David B. Dunson, National Institute of Environmental Health Sciences

Many biomedical studies collect data on times of occurrence for a health event that can occur repeatedly, such as infection, hospitalization, recurrence of disease, or tumor onset. To analyze such data, it is necessary to account for within-subject dependency in the multiple event times. Motivated by data from studies of palpable tumors, this article proposes a dynamic frailty model and Bayesian semiparametric approach to inference. The widely used shared frailty proportional hazards model is generalized to allow subject-specific frailties to change dynamically with age while also accommodating non- proportional hazards. Parametric assumptions on the frailty distribution are avoided by using Dirichlet process priors for a shared frailty and for multiplicative innovations on this frailty. By centering the semiparametric model on a conditionally-conjugate dynamic gamma model, we facilitate posterior computation and lack of fit assessments of the parametric model. Our proposed method is demonstrated using data from a cancer chemoprevention study.

email: mpennell@email.unc.edu

EXTENSIONS OF MULTIVARIATE SURVIVAL ANALYSIS TO INCLUDE GROUPED FAILURE TIME DATA WITH APPLICATION IN QUALITY OF LIFE

Denise A. Esserman*, Columbia University
Andrea B. Troxel, University of Pennsylvania School of Medicine

In oncology trials, quality of life (QOL) measures are often obtained as secondary endpoints. Especially in palliative settings, these endpoints can be just as important as survival or progression-free survival in making decisions about treatment. Many continuous QOL measures have natural thresholds that make them easily convertible into event-time data. QOL measures differ from survival measures, however, in that they are often observed at specific intervals rather than continuously. This can result in interval-censored or grouped survival data; a QOL event is defined as occurring in the interval in which the QOL measure surpasses the threshold value of the scale. We adapt methods for bivariate survival data to this setting, including a marginal model and a bivariate gamma frailty model.

email: dae2001@columbia.edu

A JOINT FRAILTY MODEL FOR SURVIVAL TIME AND GAP TIMES BETWEEN RECURRENT EVENTS

Xuelin Huang*, University of Texas M. D. Anderson Cancer Center
Lei Liu, University of Virginia School of Medicine

Patients experience disease recurrences and eventially die. It is often of interest to estimate the distributions of gap times between disease recurrences and the distribution of survival time. However, this common problem has a few challenges. The first one is that gap times are subject to dependent censoring by death. The second is that long early gap times will prevent the observation of late gap times. The empirical distributions of late gap times are biased by those subjects who have short early gap times. We show that these problems can be solved by making a very reasonable assumption, that is, gap times for the same subject share a common frailty, and this frailty affects survival. This assumption is implemented in a joint proportional hazards frailty model. The EM algorithm is used with Markov chain Monte Carlo simulation in the E-steps. We estimate the distributions of gap times and survival times, and also the effects of covariates. Simulation studies and data analysis for an real example are presented to illustrate and verify the proposed method.

email: xlhuang@mdanderson.org

A STRATIFIED FRAILTY GAP TIME MODEL FITTED BY A RE-CENSORING METHOD

Lei Liu*, University of Virginia School of Medicine
Xuelin Huang, University of Texas M. D. Anderson Cancer Center

Analysis of gap times between recurrent events has attracted increasing attention recently. In this paper, we are interested in modeling gap time data by gamma frailty proportional hazard models with stratified baseline hazards by the order of gaps. We propose the model and investigate the estimation procedure. We find that the frailty variance is often underestimated for finite sample size because of the sparseness of higher order gap times. We propose to reduce the bias by re-censoring recurrent event data by the minimum of the original censoring time and the k-th recurrent event time, where k is a pre- specified number. The new censoring time is still independent of the recurrent event process as it is the minimum of a type I and type II censoring. We demonstrate by simulation that with a good choice of k, the estimate of frailty variance behaves reasonably well and may reduce the bias from original censoring mechanism. As a good trade-off of unbiasedness and efficiency, the re- censoring method is preferable in practice.

email: liulei@virginia.edu

MULTIPLE IMPUTATION FOR NON-NORMAL MISSING DATA USING TUKEY'S gh DISTRIBUTION

Yulei He*, University of Michigan
Trivellore E. Raghunathan, University of Michigan

Multiple imputation (Rubin 1987) creates 'complete data' by 'fill-in' missing items using draws from their posterior predictive distributions. For most of existing imputation routines (such as in SAS and S-plus), the imputation algorithm is often based on multivariate normal distributional assumption. However, when data do not fit well with the normal assumption, normal-based imputation may lead to distorted distribution of 'complete data'. We propose a multiple imputation method for non- normal missing data based on Tukey' gh distribution (Tukey 1977). We have considered two settings. In the first setting, error terms of regression models are modelled using gh distribution. In the second setting, multivariate gh distribution is used for modelling multivariate non-normal data. The performance of the proposed method is evaluated through simulations. The proposed imputation approach is applied to the data set of National Health and Nutrion Examination Survey III.

email: yuleih@umich.edu

MULTIPLE IMPUTATION FOR MARGINAL AND LINEAR MIXED EFFECTS MODELS IN THE ANALYSIS OF LONGITUDINAL DATA WITH INFORMATIVE MISSINGNESS

Wei Deng*, The Ohio State University
Lei Shen, The Ohio State University

The method of multiple imputation is widely used to handle missing data. It calls for imputing draws from a predictive distribution and incorporates the sampling variability due to the missing values. However, multiple imputation for longitudinal data when missingness is not at random has not been well studied. We consider two commonly used approaches to analyze longitudinal data, mixed effects models and marginal models, under informative missingness. For the former, we apply multiple imputation using conditional models. For marginal models, which do not specify the joint distribution of the data, we generate imputed values based on linear mixed models. That is, the imputer's model differs from the analyst's model. The performance of multiple imputation is studied under a variety of circumstances.

email: deng.32@osu.edu

SMALL SAMPLE AND ASYMPTOTIC RELATIONSHIPS BETWEEN MULTIPLE IMPUTATION, MAXIMUM LIKELIHOOD, AND FULLY BAYESIAN METHODS

Qingxia Chen*, University of North Carolina at Chapel Hill
Joseph G. Ibrahim, University of North Carolina at Chapel Hill

Multiple Imputation (MI), Maximum Likelihood (ML) and Fully Bayesian (FB) methods are the three most commonly used model-based approaches in missing data problems. In this paper, we derive small sample and asymptotic expressions of the estimates and standard errors for these three methods, investigate the small and large sample properties of the estimates, and fully examine how these estimates are related for the three approaches in the linear regression model when the responses or covariates are missing at random (MAR). We show that when the responses are MAR in the linear model, the estimates of the regression coefficients using these three methods are asymptotically equivalent to the complete case (CC) estimates under very general conditions. With MAR continuous covariates in the linear model, we derive the imputation distribution under proper MI, the iterative formula of the estimates and closed form expressions for the standard errors under the ML method via the EM algorithm, as well as closed form full conditional distributions for Gibbs sampling under the FB framework. Simulations are given to compare the properties of the three methods when the responses or covariates are MAR. A real data set from a liver cancer clinical trial with one missing covariate is analyzed using the CC, MI, ML, and FB methods.

email: qchen@bios.unc.edu

ESTIMATING THE DOSE RESPONSE RELATIONSHIP FOR OCCUPATIONAL RADIATION EXPOSURE MEASURED WITH MINIMUM DETECTION LEVEL

Xiaonan (Nan) Xue*, Albert Einstein College of Medicine of Yeshiva University
Roy E. Shore, New York University School of Medicine
Mimi Y. Kim, Albert Einstein College of Medicine of Yeshiva University
Xiangyang Ye, New York University School of Medicine

Occupational exposures are often recorded as zero when the exposure is below the minimum detection level (BMDL). This can lead to an underestimation of the doses received by individuals and can lead to biased estimates of risk in occupational epidemiologic studies. The extent of the exposure underestimation is increased with the magnitude of the minimum detection level (MDL) and the frequency of monitoring. This paper uses multiple imputation methods to impute values for the missing doses due to BMDL. A Gibbs sampling algorithm is developed to implement the method, which is applied to two distinct scenarios: when dose information is available for each measurement (but BMDL is recorded as zero or some other arbitrary value), or when the dose information available represents the summation of a series of measurements (e.g., only yearly cumulative exposure is available but based on, say, weekly measurements). Then the average of the multiple imputed exposure realizations for each individual is used to obtain an unbiased estimate of the relative risk associated with exposure. Simulation studies are used to evaluate the performance of the estimators. As an illustration, the method is applied to a sample of historical occupational radiation exposure data from the Oak Ridge National Laboratory.

email: xxue@aecom.yu.edu

88. GENE EXPRESSION ANALYSIS IN CANCER RESEARCH

A COMPARISON OF GENE EXPRESSION MEASUREMENTS FROM COMMERCIAL MICROARRAY PLATFORMS

Karla V. Ballman*, Mayo Clinic College of Medicine
Christopher P. Kolbert, Mayo Clinic College of Medicine
Sreekumar Raghavakaimal, Mayo Clinic College of Medicine

Numerous commercial microarray platforms for measuring gene expression levels are currently available. The formats of these systems differ in terms of material used (short oligonucleotides, long oligonucleotides, or cDNA) and number of samples per array (single-channel or two- channel). We performed a study to compare gene expression measurements of identical RNA preparations obtained from four commercially available microarray platforms. Two technical replicates were prepared for RNA collected from two different prostate cancer cell lines (PC3 and DU145). RNA was labeled and hybridized to the microarrays according to manufacturers' protocols. Gene expression values were obtained using each platform's standard software. Gene expression values were also obtained using our preferred analytic methods. The goal was to assess reproducibility within each platform as well as the correlation of results among the platforms. We also compared results between the two different methods used to generate gene expression values. None of the platforms appeared to be clearly superior in terms of reproducibility using gene expression values generated by the platform's software. We also found the correlation of results among the platforms to be relatively poor. Finally, the choice of methodology for generating gene expression values greatly influenced the results.

email: ballman@mayo.edu

A TWO-STAGE MIXTURE MODEL STRATEGY FOR META-ANALYSIS OF MICROARRAY DATA

Ronglai Shen*, University of Michigan
Debashis Ghosh, University of Michigan
Arul M. Chinnaiyan, University of Michigan

An increasing number of studies have profiled tumor specimens using distinct microarray platforms and analysis techniques. With the accumulating amount of microarray data, one of the most challenging tasks is to develop robust statistical models to integrate the findings. In this study, we propose a two-stage approach for meta-analysis of data from microarrays using a mixture model-based data transformation technique. The methods are applied to validate and integrate independent findings from four microarray studies in breast cancer. Comparisons with univariate analysis methods and study-specific global standardization are also examined.

email: rlshen@umich.edu

BAYESIAN COVARIANCE SELECTION

Adrian Dobra*, Duke University

We present a novel structural learning method called HdBCS that performs covariance selection in a Bayesian framework for datasets with tens of thousands of variables. HdBCS is based on the intrinsic connection between graphical models on undirected graphs and graphical models on directed acyclic graphs (Bayesian networks). We show how to produce and explore the corresponding association networks by Bayesian model averaging across the models identified. We illustrate the use of HdBCS with an example from a large- scale gene expression study of breast cancer.

email: adobra@stat.duke.edu

MODEL SELECTION TECHNIQUES IN GENE EXPRESSION PROFILING FOR PREDICTING BREAST CANCER OUTCOME

Zhaoling Meng*, Merck & Co., Inc
Bret Musser, Merck & Co., Inc.

Van't Veer et. al. (2002) constructed a 70-gene classifier for predicting breast cancer survival based 78 patients. In an independent 19 patients validation set, 1 out of 12 (8%) poor prognostic and 1 out of 7 (14%) good prognostic patients were misclassified. Although holding encouraging prediction power, a 70-gene classifier could be viewed as a black-box since it is hard to fully understand and follow up each of the 70 genes biologically. We proposed two methods to reduce gene number in the classifier without sacrificing power; furthermore, the impact of pre- filtering genes on the predictive ability of the modeling techniques was also assessed. A spectral decomposition method was applied based on gene expression correlation; a 14-gene classifier built misclassified 1 of 12 (8%) poor and 2 out of 7 (29%) good prognostic patients when starting with after pre-filtering genes; a 22-gene classifier built misclassified 0 of 12 poor and 1 out of 7 (14%) good prognostic patients when starting with all genes without pre-filtering. Also, logistic regression was used in a forward step-wise selection fashion using pseudo-cross-validation; a model containing only 5 genes misclassified 2 out of 7 (29%) good prognostic and 1 out of 12 (8%) poor prognostic patients.

email: zhaolingm@yahoo.com

ANALYSIS OF GENE EXPRESSION DATA USING SPLIT-MERGE MARKOV CHAIN MONTE CARLO

Sonia Jain*, University of California, San Diego
Radford M. Neal, University of Toronto

The inferential problem of associating data in high dimensions to mixture components is difficult when components are nearby or overlapping. We introduce a new split-merge Markov chain Monte Carlo technique that efficiently classifies observations by splitting and merging mixture components of a nonconjugate Bayesian mixture model. Our method, which is a Metropolis-Hastings procedure with split-merge proposals, samples clusters of observations simultaneously rather than incrementally assigning observations to mixture components. Split-merge moves are produced by exploiting properties of a restricted Gibbs sampling scan. We apply our split-merge technique to a cancer classification problem, in which patients are clustered according to leukemia type based on their gene expression data obtained from DNA microarray experiments.

email: sojain@ucsd.edu

CONSTRUCTING PROGNOSTIC GENE SIGNATURES FOR CANCER SURVIVAL

Derick R. Peterson*, University of Rochester Medical Center

Modern micro-array technologies allow us to simultaneously measure the expressions of a huge number of genes, some of which are likely to be associated with cancer survival. While such gene expressions are unlikely to ever completely replace important clinical covariates, evidence is already beginning to mount that they can provide significant additional predictive information. The difficult task is to search among an enormous number of potential predictors and to correctly identify most of the important ones, without mistakenly identifying too many spurious associations. Many commonly used screening procedures unfortunately over-fit the training data, leading to subsets of selected genes that are unrelated to survival in the target population, despite appearing associated with the outcome in the particular sample of data used for subset selection. And some genes might only be useful when used in concert with certain other genes and/or with clinical covariates, yet most available screening methods are inherently univariate in nature, based only on the marginal associations between each predictor and the outcome. While it is impossible to simultaneously adjust for a huge number of predictors in an unconstrained way, we propose a method that offers a middle ground where some partial adjustments can be made in an adaptive way, regardless of the number of candidate predictors.

email: peterson@bst.rochester.edu

89. BAYESIAN AND NON-BAYESIAN APPROACHES TO COMPETING RISKS

A MARGINAL CONDITIONAL MODEL FOR MULTIVARIATE SURVIVAL DATA

Glen A. Satten*, Centers for Disease Control and Prevention
Somnath Datta, University of Georgia

A commonly used multivariate survival model is the marginal analysis proposed by Lin, Wei and Weissfeld (1989 JASA 84:1065-1073), in which a (marginal) Cox model is proposed for each of the survival times. These marginal models are fit separately, and a subsequent adjustment is made to the covariances of the parameter estimates to account for the fact that the marginal models are fit using possibly dependent data. Although the marginal model is easy to implement, it has the disadvantage that in some cases the parameter estimates are difficult to interpret when competing risks act to make some persons not at risk for certain events. For example, in a marginal analysis of risk factors for a 3rd heart attack, a Cox model would be proposed for time to the third heart attack. Persons who died before their 2nd heart attack would be considered censored, even though they are clearly not at risk for a 3rd heart attack. Here, we propose a marginal conditional model that uses a sample- reweighting scheme to fit a marginal Cox model for each failure time of interest, conditional on being at risk for the event of interest.

e-mail: gsatten@cdc.gov

BAYESIAN ANALYSIS OF COMPETING RISKS IN CANCER SURVIVAL

Sanjib Basu*, Northern Illinois University and Rush University Medical Center

Survival data from cancer studies can be viewed as resulting from the competing risks of cancer and ‘other causes’ (collection of all other risks). A popular approach for analysis of such data is the relative survival approach. Alternative approaches include the net survival approach based on latent, potential survival times, and the crude or cause-specific model, which emphasizes observed (non-latent) quantities. Statistical analysis of such data is often further complicated by partial masking when the cause of death is not known for a subgroup of patients. We discuss these different approaches and describe Bayesian parametric and semiparametric competing risks analyses of cancer survival data.

e-mail: basu@niu.edu

JOINT COMPETING RISK MODELING FOR ASSESSING IMPORTANT PSA MARKERS IN PREDICTING PROSTATE CANCER SPECIFIC MORTALITY

Ming-Hui Chen*, University of Connecticut
Joseph G. Ibrahim, University of North Carolina at Chapel Hill
Anthony V. D'Amico, Harvard University

In this paper, we develop a joint competing risk models for longitudinal PSA and prostate cancer survival data. We investigate whether several important PSA markers recently identified in the prostate cancer literature, such as the PSA velocity and PSA doubling time, are significant prognostic factors or surrogate end points for prostate cancer specific mortality under the joint competing risk modeling. Novel EM and Monte Carlo sampling algorithms are developed for carrying out statistical inference. A study cohort formed from multi-institutional databases containing baseline, treatment, and follow up information on men treated with surgery or radiation for clinical stage T1c- 4NxMo prostate cancer is used to illustrate the proposed methodology.

e-mail: mhchen@stat.uconn.edu

90. INTEGRATING MULTIPLE SOURCES OF GENOMIC DATA

STATISTICAL METHODS FOR ChIP-Chip HIGH-DENSITY OLIGONUCLEOTIDE ARRAY DATA

Sunduz Keles*, University of Wisconsin–Madison
Mark J. van der Laan, University of California, Berkeley
Sandrine Dudoit, University of California, Berkeley
Simon E. Cawley, Affymetrix

Cawley et al. (2004) have recently mapped the locations of binding sites for three transcription factors along human chromosomes 21 and 22 using ChIP-Chip experiments. ChIP-Chip experiments are a new approach to the genome-wide identification of transcription factor binding sites and consist of chromatin (Ch) immunoprecipitation (IP) of transcription factor-bound genomic DNA followed by high density oligonucleotide hybridization (Chip) of the IP-enriched DNA. We investigate the ChIP-Chip data structure and propose novel statistical methods for inferring the location of transcription factor binding sites from these data. The proposed methods involve testing for each probe whether it is part of a bound sequence or not using a scan statistic that takes into account the spatial structure of the data. Different multiple testing procedures are considered for controlling the family-wise error rate and false discovery rate. Application of the proposed methods to ChIP-Chip data for transcription factor p53 identified many potential target binding regions along human chromosomes 21 and 22. Among these identified regions, 18% fall within a 3kb vicinity of the 5'UTR of a known gene or CpG island, 31% fall between the codon start site and the codon end site of a known gene but not inside an exon. More than half of these potential target sequences contain the p53 consensus binding site or very close matches to it. Moreover, these target segments include the 13 experimentally verified p53 binding regions of Cawley et al. (2004), as well as 49 additional regions that show higher hybridization signal than these 13 experimentally verified regions.

e-mail: keles@stat.wisc.edu

IMPROVING THE ACCURACY OF PROTEIN-PROTEIN INTERACTION NETWORK USING LOCAL GRAPH STRUCTURE AND COMPARATIVE GENOMICS

Peter J. Park*, Children's Hospital Boston
Jung-Ah Lim, Children's Hospital Boston

Several large scale protein-protein interaction maps have been developed recently. However, these maps contain a high rate of false positives due to various shortcomings of the experimental procedures. We will describe our efforts to estimate more accurate probabilities of interactions between proteins, using the structure of the interaction network as well as by integrating data from different organisms.

e-mail: peter_park@harvard.edu

PATHWAY-BASED ANALYSIS OF DNA MICROARRAY DATA

Marina Vannucci*, Texas A&M University

Classification has received a lot of attention in DNA microarray data analysis. Current methods build classifiers without using prior knowledge on gene function. Here, we propose a method to analyze gene expression data in the context of known biological pathways. We use the Kyoto Encyclopedia of Genes and Genomes (KEGG) database to extract the pathway membership of each gene. Our approach addresses simultaneously the problems of classifying samples and identifying important genes. We illustrate the methodology using recently published gene expression data from a breast cancer study, where the interest is in isolating gene expression signatures that are predictive of a short time interval to distant metastases.

e-mail: mvannucci@stat.tamu.edu

APPROACHES FOR REDUCING SPURIOUS CLUSTER IDENTIFICATION IN SCAN STATISTICS

Howard S. Burkom*, Johns Hopkins University

The public health community has devoted much energy to collecting a variety of data sources for biosurveillance. The spatiotemporal scan statistic has gained attention because the Kulldorff search paradigm provides the location and extent of anomalous case clusters along with a measure of the significance of each cluster. This approach also avoids preselection bias— insofar as the spatial data resolution allows—and controls for multiple testing. However, several issues complicate the adaptation of this approach from classical epidemiology designs to prospective monitoring. The data to be clustered may exhibit temporal behavior that varies by subregion. Estimating local rates may be difficult because underlying population counts may be unavailable for reasons of data privacy or ownership. For some data sets, the data histories of the respective subregions are the only information available for estimating this distribution. This presentation considers approaches for using random- effects modeling to combine the data histories of the various subregions to estimate the multinomial spatial case distribution. The reduction in unwanted cluster identifications is expressed as a function of the improvement in the estimate of this distribution. Separate findings for large-count and small-count data streams are compared.

e-mail: Howard.Burkom@jhuapl.edu

BAYESIAN SPATIAL SURVEILLANCE OF SMALL AREA DISEASE EVENTS: PARTICLE FILTERING METHODS

Carmen L. Vidal Rodeiro, University of South Carolina
Andrew B. Lawson*, University of South Carolina

The analysis of small area health data online in a surveillance context is of considerable current interest given the need for enhanced biosurveillance. Many approaches can be adopted to this problem. In this talk we examine the use of computational algorithms for particle filtration to allow fast updating. We consider a reasonably sophisticated spatial and spatio- temporal Hierarchical Bayesian model. A comparison of computational efficiency is made and an application of the S-T model to the New England ER addmissions data of Kleinman is made.

e-mail: luciavidal@yahoo.com

BIOSURVEILLANCE GEOINFORMATICS OF HOTSPOT DETECTION AND PRIORITIZATION FOR BIOSECURITY

G. P. Patil, Pennsylvania State University
Stephen L. Rathbun, Pennsylvania State University

Hotspot means something unusual—an anomaly, aberration, outbreak, elevated cluster, critical resource area, etc. The responsible factors may be natural, accidental, or intentional. This presentation describes the upper level set (ULS) scan statistic for hotspot detection across geographic regions and across networks. The method is computationally efficient and can identify clusters of arbitrary shape including those that may not be adequately captured by the traditional circle-based scan statistic. The ULS statistic extends to the space-time domain where, because it allows for arbitrarily shaped space-time hotspots, it can characterize the temporal evolution of a spatial hotspot. This leads to the "typology of space-time hotspots." Changing patterns of urban poverty in different metropolitan areas of the U.S. are examples of such typologies. We also describe methods for multi-criteria prioritization of identified hotspots, employing the notion of linear extensions of partially ordered sets (poset). Except for toy examples, complete enumeration of all linear extensions of a poset is impossible so we employ MCMC sampling with a specified (typically uniform) target distribution.

e-mail: taillie@stat.psu.edu

---

92. RISK RANKING AND DISEASE MAPPING

TYPE S AND TYPE M ERROR RATES FOR RANKING COMPARISONS

Samantha R. Cook*, Columbia University
Andrew Gelman, Columbia University
Francis Tuerlinckx, University of Leuven

Disease ranking often involves pairwise comparisons, which are calibrated using the Type 1 and Type 2 error rates. Claims of a true difference being positive or negative are made with confidence if the 95% interval for the difference excludes zero. We present two additional types of errors that may be more relevant in this setting than the Type 1 and Type 2 error rates. A Type S (for sign) error occurs when you claim with confidence that the true difference is positive and the difference is in fact negative, or vice versa. A Type M (for magnitude) error occurs when a claim is made with confidence and the magnitude of the true difference is smaller than the magnitude of the values in the 95% interval. We compute Type S and Type M error rates for Bayesian and frequentist procedures, focusing on the hierarchical normal model. We also present results for multiple comparison situations.

e-mail: cook@stat.columbia.edu

OPTIMAL SURVIVAL CURVE RANKING (OSCR): APPLICATION TO ADJUSTMENT OF AIDS REPORTING DELAY

Vanja Dukic*, University of Chicago
Peter Bouman, Northwestern University
Xiao-Li Meng, Harvard University

Time delay between a new AIDS diagnosis and its report to the CDC, historically ranging between a couple of weeks and a couple of years, presents a significant problem when trying to predict future AIDS incidence and health care burden. The reporting delay needs to be correctly estimated and adjusted for in order to avoid potentially serious downward bias. We examine case reports from 39 large US cities, received by the CDC as of the end of December 2001 and published in the APIDS database. We employ Bayesian multi-resolution methodology to estimate city-specific hazards of reporting delay, adjusting for patient covariates and within-city correlation. We describe the ranking of the 39 US cities according to their reporting delay distributions based on the optimal survival curve ranking (OSCR) procedure. We discuss uncertainty in the reporting delay estimates and in the resulting ranking, and present a graphical approach to visualize this uncertainty.

e-mail: vanja@uchicago.edu

BAYESIAN RANKING METHODS WITH APPLICATIONS TO DISEASE MAPPING

Thomas A. Louis*, Johns Hopkins University
Rongheng Lin, Johns Hopkins University
Susan M. Paddock, Rand Corporation
Greg Ridgeway, Rand Corporation

Prioritizing environmental assessments, health services evaluations, school effectiveness studies and identification of differentially expressed genes depend on the relative position (ranks) of unit-specific attributes. Invalid ranks or inappropriate interpretation can have serious health, financial, policy and scientific consequences. When estimation uncertainty varies over units, ranks produced from hypothesis test statistics inappropriately identify units with relatively low variance as extreme because these tests have highest power; ranks produced from the MLEs inappropriately identify units with relatively high variance as extreme. Therefore, effective ranking depends on properly accommodating both signal and noise and Bayesian modeling coupled with loss functions provides the necessary guidance. We compare performance of several loss function based ranking methods, MLE-based ranks, ranks based on posterior means and hypothesis-test based ranks. We report simulation-based and data-analytic measures of performance; we identify general issues and trade-offs in choosing among methods. Performance evaluations show that in many applications even optimal ranks perform poorly and an uncertainty assessment is essential. We illustrate the foregoing with disease mapping, provider profiling and microarray analysis examples.

e-mail: tlouis@jhsph.edu

93. TO MIX OR NOT TO MIX

INTERPRETATION OF A MIXTURE MODEL

Bruce G. Lindsay*, Pennsylvania State University
Surajit Ray, Statistical and Applied Mathematical Sciences Institute (SAMSI)

A finite mixture model creates a decomposition of a population density into a set of homogeneous components. One popular model is based on using the multivariate normal distribution for the components. One interpretation of the estimators in this setting is that the weights given to the components, together with their parameter values, provide a description of the heterogeneous features of the data, while the multivariate normal components represent the residual variability as homogeneous 'white noise'. Our theoretical examination of the topographical features of mixtures of multivariate normal densities shows that the underlying structure can be considerably more complex than this. For a mixture of two components, for example, there is a one dimensional curve through the sample space that contains all the maxima, minima, and saddlepoints of the density, for any weights on the two components. This mathematical simplification can then be used to show how complex the density surface can be. Our theoretical tools lead to some new ideas about how to describe data with fitted mixture models.

e-mail: bgl@psu.edu

REMARKS ON MIXTURES OF REGRESSIONS

David W. Scott*, Rice University

Robust regression techniques can help with messy data contaminated by outliers. A related problem occurs when the data represent a mixture of regressions, perhaps with significant overlap. We examine how kernel methods, minimum distance parametric modeling, and mixture models may be used to address such data.

e-mail: scottdw@rice.edu

LOCAL LIKELIHOODS VERSUS LOCAL MIXTURE LIKELIHOODS

Ramani S. Pilla*, Case Western Reserve University
Catherine Loader, Case Western Reserve University

Local likelihood, introduced by Tibshirani and Hastie (1987 JASA, 559-567), is an extension of local regression for analyzing a variety of data structures, including point processes, count, survival and binary data. Extensive theoretical and methodological advances have been made since then; however, largely for cases where there is a single function, such as the conditional mean, to estimate. This research develops inferential theory for extending the local likelihood method for responses from a mixture model. Such an interplay enables further generalizations of the error structure, such as multimodal and skewed distributional shapes as well as providing a natural interpretation for overdispersion. However, this also leads to several challenges when locations, mixing weights and the number of components are allowed to vary spatially throughout the predictor space. The question of whether it is advantageous to apply mixture models in the context of local likelihood for model building will be addressed.

e-mail: pilla@case.edu

---

94. STATE SPACE MODELS AND TIME SERIES ANALYSIS

STATE SPACE MODELS OF IMMUNE RESPONSES UNDER TREATMENT IN PLASMA AND LYMPH NODES

Wai-Yuan Tan, University of Memphis
Ping Zhang*, Middle Tennessee State University
Xiaoping Xiong, St. Jude Research Hospital

It is well documented that in many cases, most of free HIV are generated in the lymphoid tissues rather than in the plasma.To assess effects of drugs, in this paper we have thus developed a state space model for HIV pathogenesis involving both plasma and lymph nodes and flow of HIV from lymph nodes to plasma. We have applied this model and the theory to the data in Lafuillade et al(1996), in which RNA virus copies per mm^3 were observed in both plasma and lymph nodes at different time points since treatment. Our results showed that the HIV infection process in the lymph nodes was about 10^2 ~ 10^3 times faster than that in the plasma, indicating that more than 99% of HIV are produced in the lymph nodes. This analysis provides strong support for the model by Bajaria et al. (2000, 2002) who have assumed that the HIV is produced in the lymph nodels more than in the plasma.

email: pingzhang0407@yahoo.com

RANDOM COEFFICIENT TRANSFER FUNCTION MODEL FOR PANEL DATA

Hyunyoung Choi*, University of Illinois at Urbana–Champaign
Hernando Ombao, University of Illinois at Urbana–Champaign
Bonnie Ray, IBM T. J. Watson Research Center

We present a random coefficient transfer function model applicable to a panel of time series data. The random coefficient transfer function model allows for individual subject variation in response to an input series, while assuming a common distribution for the individual effects. Additionally, the output series are each assumed to follow ARMA models with commonly distributed parameters across series. The model is estimated in the Bayesian framework using Gibbs sampling, which also allows for handling of missing data within each series. Through simulation, we compare the results of fitting the proposed model to those obtained from fitting univariate transfer function models or panel transfer function models with deterministic, common input effects. We then present an application to monitoring electroencephalogram signals for a set of subjects to determine the overall impact of a common stimulus.

email: hchoi9@uiuc.edu

INVESTIGATION OF SYNCHRONY BETWEEN BRAIN REGIONS USING WAVELET COHERENCE

Bing Gao*, University of Illinois at Urbana-Champaign
Hernando Ombao, University of Illinois at Urbana-Champaign
Christopher Edgar, University of Illinois at Urbana-Champaign

The goal in this paper is to investigate how synchrony between brain regions varies according to different stimuli in a Stroop task experiment. The data in our analysis consists of brain wave recordings (ERPs) at 4 regions (2 in the frontal and 2 in the posterior areas) of 10 healthy normal subjects. The proposed measure of synchrony is wavelet coherence which is defined as the correlation between the wavelet-packets filtered output signals. Due to the inherent non-stationarity in the ERPs, our approach uses a moving window in order to obtain instantaneous wavelet coherence. Unlike Fourier-based coherence, wavelet coherence describes the correlation on multiple scales, and can supply both temporal and frequency information. Our analysis shows that wavelet coherence within the frontal and right area is significantly different between congruent and incongruent tasks at the beta frequency band of 16-24Hz at about 300-800 ms from stimulus presentation. Moreover, our method also shows that there is a significant correlation between reaction time and the time to maximal coherence within the frontal area for the incongruent case.

email: binggao@uiuc.edu

AUTOMATED PEAK IDENTIFICATION IN A TIME-OF-FLIGHT SPECTRUM

Haijian Chen*, College of William and Mary
Eugene R. Tracy, College of William and Mary
William E. Cooke, College of William and Mary
Michael W. Trosset, College of William and Mary

Although mass spectrometry has become one of the standard tools of proteomic research, most time-of-flight (TOF) spectrometers produce very large raw data sets that must be preprocessed to identify the mass peaks corresponding to important biological molecules. Under survey conditions, where the positions of the desired mass peaks are not known beforehand, a TOF instrument requires a peak-picking procedure to distinguish mass peaks from a slowly varying background. We have developed and automated a peak picking algorithm based on a maximum likelihood approach that effectively and efficiently detects peaks in a time-of-flight (TOF) spectrum. This approach produces maximum likelihood estimates of peak positions and amplitudes, and simultaneously develops estimates of the uncertainties in each of these quantities. This peak picking algorithm was developed for TOF-SIMS spectra but we are currently extending it to SELDI spectra, even thought the noise characteristics are completely different for the two technologies. We will show that even when adjacent peaks strongly overlap, this technique identifies the constituent peaks, and also provides additional tagging information to mark overlapping peaks as possible problem cases requiring further investigation.

email: hxche2@wm.edu

VARIANCE-COVARIANCE ESTIMATION WITH AN APPLICATION TO INTERCELLULAR SIGNALING

Scott H. Holan*, University of Missouri–Columbia

In multicellular organisms cells constantly exchange signals. This intercellular signaling is crucial for the coordination and regulation of cell behaviors. To determine if paracrine intercellular signaling is important in determining characteristics of colonocytes, serial sections of rodent colon tissue are examined using fluorescence microscopy. Colon architecture coupled with the method of tissue collection produce arrays of cells with a one-dimensional spatial orientation. Therefore, to evaluate the correlation between neighboring cells we use multiple time series methodology. More specifically, we assume the process is covariance stationary and use maximum likelihood estimation to estimate the process variance-covariance matrix. Additionally, we form a likelihood ratio test for determining whether or not neighboring cells exhibit autocorrelation. Lastly, we demonstrate the proposed methodology using simulation as well as a real data analysis.

email: holans@missouri.edu

NEUROPATHOLOGIC MEDIATORS OF THE ASSOCIATION BETWEEN THE APOLIPOPROTEIN E EPSILON4 ALLELE AND CLINICAL DEMENTIA

Yan Li*, Rush University Medical Center
Julia L. Bienias, Rush University Medical Center
David A. Bennett, Rush University Medical Center

Dementia is due to the accumulation of several neuropathologic indices in the brain. Much has been discovered in recent years regarding risk factors for dementia, but the underlying neuropathological mediators that link risk factors to dementia are still far from understood. The presence of one or two copies of the å4 allele of the apolipoprotein E gene (ApoE å4) has been shown to be a predictor of clinical dementia. Because the presence of the allele precedes the phenotypic expression of accumulation of brain pathology and of dementia, we undertook to examine which neuropathological indices were in the causal chain between ApoE å4 and dementia. Using the approach of statistical mediation, we tested whether different neuropathologic indices of Alzheimer's disease (neuritic plaques and neurofibrillary tangles), measured post-mortem, changed the association of ApoE å4 to the clinical diagnosis of dementia proximate to death. We also investigated whether the presence of cerebral infarctions, measured post-mortem, mediated the association between ApoE å4 and dementia, as both are known to be associated with both ApoE allele status and dementia in separate models. Research supported by National Institute on Aging grants P30 AG10161, R01 15819, and R01 17917.

email: yan_li@rush.edu

ARTIFICIAL CENSORING FOR RANDOMIZED CLINICAL TRIALS IN THE PRESENCE OF NON-RANDOM NON-COMPLIANCE

Long-Long Gao*, University of Pennsylvania School of Medicine
Marshall M. Joffe, University of Pennsylvania School of Medicine

Randomized clinical trials often have non-compliance with the assigned treatment, which poses a challenge for correctly estimating treatment effects. G-estimation exploits the randomization assumption to estimate the causal effect of treatment received, in contrast to intention-to-treat (ITT) which ignores non-compliance and in contrast to as-treated analysis which ignores randomization. However, sometimes low or high values of the outcome variable can not be measured, and so the outcome data are said to be censored. For example, the glomerular filtration rate (GFR), a measure of renal function, is difficult to measure for subjects who are on hemodialysis. We show how to deal with this censoring in the presence of non-compliance by proposing an artificial censoring method to overcome this obstacle. Under the artificial censoring, we treat some subjects as censored even though their outcomes are actually observed. We apply this artificial censoring method to the randomized clinical trial of Modification of Diet in Renal Disease (MDRD).

email: lgao@cceb.upenn.edu

CAUSAL INFERENCE IN HYBRID INTERVENTION TRIALS INVOLVING TREATMENT CHOICE

Qi Long*, University of Michigan
Roderick J. A. Little, University of Michigan
Xihong Lin, University of Michigan

Randomized allocation of treatments is a cornerstone of experimental design, but has drawbacks when a limited set of individuals are willing to be randomized, or the act of randomization undermines the success of the treatment. Choice-based experimental designs allow a subset of the participants to choose their treatments. We discuss here causal inferences for experimental designs where some participants are randomly allocated to treatments and others receive their treatment preference. This paper was motivated by the ``Women Take Pride'' (WTP) study (Janevic et al., 2001), a doubly randomized preference trial (DRPT) to assess behavioral interventions for women with heart disease. We propose a model that allows us to estimate the causal effects in the subpopulations defined by treatment preferences and the preference effects for a DRPT, and develop an EM algorithm to compute maximum likelihood estimates of the model parameters. The method is illustrated by analyzing treatment compliance of the WTP data. Our results show that there were strong preference effects in the WTP study, that is, women assigned to their preferred treatment were more likely to comply. We also expand these methods to handle a broader class of designs, and discuss alternative designs from the perspective of the strength of assumptions required to make causal inferences.

email: qlong@umich.edu

POLYDESIGNS IN CAUSAL INFERENCE

Fan Li*, Johns Hopkins University
Constantine E. Frangakis, Johns Hopkins University

In an increasingly common class of studies, the goal is to evaluate causal effects of treatments that are only partially controlled by the investigator. In such studies there are two conflicting features: (1) a model on the full cohort design and data can identify the causal effects of interest, but can be sensitive to extreme regions of that design's data, where model specification can have more impact; and (2) models on a reduced design (i.e., a subset of the full data), e.g., conditional likelihood on matched subsets of data, can avoid such sensitivity, but do not generally identify the causal effects. We propose a framework to assess how inference is sensitive to designs by exploring combinations of both the full and reduced designs. We show that using such a ``polydesign'' framework generates a rich class of methods that can identify causal effects and that can also be more robust to model specification than methods using only the full design. We discuss implementation of polydesign methods, and provide an illustration in the evaluation of a Needle Exchange Program.

email: fli@jhsph.edu

A FORMAL APPROACH FOR DEFINING AND IDENTIFYING THE FUNDAMENTAL EFFECTS OF EXPOSURES ON DISEASE FROM SETS OF EXPERIMENTS CONDUCTED ON POPULATIONS OF NON-IDENTICAL SUBJECTS

Steven D. Mark*, National Cancer Institute

Causal inference is the branch of statistics that proposes formal systems to bridge the gap between 'association' and 'causation.' Rubin's causal model, often called the counterfactual model, provides the mathematical framework for nearly all such inference in epidemology. Fundamental to such inference is the concept that the effect of exposures could be determined by comparing groups that are identical in all ways except for the exposures. However, many questions in the study of chronic diseases in humans do not fall into that paradigm. One obvious example occurs when experiments conducted on mice are used to make inferences about the causes of cancer in humans. In this paper we propose a formal system of inference that bridges the 'association/causation gap' when the goal is to make inferences from experiments on such 'non-identical' populations. The formal inferential structure has implications for the design and analysis of a sequence of experiments that are informative in terms of evaluating both the underlying assumptions, and the hypotheses about the exposure-disease relationships of interest.

email: smark@exchange.nih.gov

BOUNDS ON CAUSAL EFFECTS IN THREE-ARM TRIALS WITH NONCOMPLIANCE

Jing Cheng*, University of Pennsylvania
Dylan S. Small, University of Pennsylvania

Three-arm trials are common in practice. In these trials, the average causal effects of treatments within principal strata are of interest for several reasons discussed in the paper. Unfortunately, even with usual assumptions, the average causal effects of treatments within principal strata are not point-identified. However, the observable data does provide useful information on the bounds of the identification regions of the parameters of interest. Under two sets of assumptions, we derive sharp bounds for the causal effects within principal strata, and construct confidence intervals to cover the identification regions. The methods are illustrated by an analysis of data from a randomized study of treatments for alcohol dependence. keywords: Three-arm trials; noncompliance; causal effect; principal strata; bounds; confidence interval.

email: jcheng@cceb.upenn.edu

A DISTRIBUTIONAL APPROACH FOR CAUSAL INFERENCE USING THE PROPENSITY SCORE

Zhiqiang Tan*, Johns Hopkins University

Drawing inferences about the effects of exposures or treatments is a common challenge in many scientific fields. We propose two new methods serving complementary purposes in causal inference. One can be used to estimate average causal effects, assuming 'no confounding' given measured covariates. The other can be used to assess the sensitivity of the estimates to possible departures from 'no confounding'. We establish asymptotic results for the methods, and also address practical issues in planning data analysis, checking propensity score models, and interpreting sensitivity parameters. Both methods are developed from a nonparametric likelihood perspective. We illustrate the methods by analyzing the data from an observational study on right heart catheterization. The talk will be based on the working paper, 'Efficient and Robust Causal Inference: A Distributional Approach,' available at http://www.bepress.com/jhubiostat/paper48.

email: ztan@jhsph.edu

---

96. CLINICAL TRIALS II

MULTI-CENTER CLINICAL TRIALS: RANDOMIZATION AND ANCILLARY STATISTICS

Lu Zheng*, Harvard University
Marvin Zelen, Harvard University

An important class of planned experiments is the multi-center randomized clinical trial. A design based analysis would rely on the permutation distribution generated by the randomization process. Ordinarily the number of patients assigned to each treatment within a center is a random variable, but is also an ancillary statistic. Another feature of multi-center randomized trials is the use of permuted blocks to allocate the treatments. The permuted blocks also generate ancillary statistics. An important principal in frequentist inference is to condition on the ancillary. Finding the exact distribution of the appropriate test statistic under these circumstances may be difficult, if not impossible. As a result we have developed an approximation to this distribution. Simulations show that the approximation works well. We have investigated the power in the context of multi-center trials with variation between institutions. Our investigations indicate that there is an increase in power, conditioning on the ancillary statistics, compared to ignoring the ancillary statistics. The increase in power is a function of the variation amongst the treatment sample sizes within institutions and may be considerable if there is large variation between institutions. The methods have been extended to group sequential trials with similar increases in power.

email: lzheng@hsph.harvard.edu

DESIGNS FOR PHASE I CLINICAL TRIALS WITH CONTINUOUS/MULTINOMIAL TOXICITIES

Zhilong Yuan*, University of Wisconsin–Madison
Rick Chappell, University of Wisconsin–Madison

Background: In phase I cancer clinical trials, adjustment for patient differences in toxicity susceptibility can be carried out with stratification into risk groups. Separate trials conducted for each risk group can lead to conflicting decisions, in which higher doses are recommended for higher risk groups. Designs with covariate-adjustment often require assumptions that clinicians may be uncomfortable with. Methods: We extend up-and-down designs, isotonic designs and the continual reassessment method (CRM) to multiple risk groups with two-way isotonic regression. The only assumption about the groups is that they can be ordered according to their toxicity risk. Results: Simulations were based on an ongoing helical tomotherapy trial. Seven different toxicity scenarios were considered. The proposed methods compared favorably to a covariate adjusted CRM. The extended up-and-down designs inherited the conservativeness from the original designs. Conclusion: Our experience demonstrates that the escalation rules of multiple risk groups can be linked, without a parametric assumption about the group-toxicity curve, to borrow strength and to ensure nonconflicting dosage recommendations.

email: zhilong@stat.wisc.edu

SAMPLE SIZE CALCULATION IN SURVIVAL TRIALS ACCOUNTING FOR TIME-DEPENDENT DYNAMICS OF NONCOMPLIANCE AND RISK

Bingbing Li*, School of Public Health, University of Minnesota
Patricia Grambsch, School of Public Health, University of Minnesota

Most of the existing methods of sample size calculations for survival trials adjust the estimated outcome event rates for noncompliance based on the assumption that noncompliance is independent of the endpoint risk although there has been published evidence that noncompliers are often at a higher risk than compliers. More recent works have started to consider the situations of informative noncompliance and different risks for noncompliers. However, the possibility of a time-varying association between noncompliance and risk has been ignored. Our analysis indicates a strong time-varying relationship between permanent withdrawal from study treatments and endpoint risk in the CONVINCE trial and we believe that this phenomena is not uncommon in survival trials. In this paper, we introduce a method of sample size calculation which can account for various assumptions about noncompliance and risk. The method is based on Lakatos (1988, Biometrics 44, 229-241) Markov models. Results with our method show that sample size can vary dramatically with different assumptions about noncompliance and risk. Power can be seriously reduced if the assumed association does not agree with the real situation.

email: bingbinl@biostat.umn.edu

MULTI-CENTER TRIALS WITH BINARY RESPONSE

Vladimir Dragalin, GlaxoSmithKline
Valerii Fedorov*, GlaxoSmithKline

Great progress has been made recently in the design and analysis of multicentre clinical trials. Different models for data have been considered, the combined response to treatment was proposed to define the overall treatment in such trials and corresponding estimators derived and analyzed. The vast majority of these results is for the continuous response case. However, there are many multicentre clinical trials in which the response is binary. The popular approach in this setting is the generalized linear mixed model. In this paper, we propose a correlated beta-binomial model for the binary response in muticentre trials. The likelihood function in this case has a closed-form and we avoid numerous multivariate numerical integrations in determining the maximum likelihood estimator (MLE). Moreover, we derive the asymptotic variance-covariance matrix of the MLE that forms the basis for optimal balancing of the number and size of centres. As an alternative to the MLE, we consider a simpler quasi- likelihood estimator. In both cases, we obtain relatively simple formulae that relate the number of centres, the total number of patients and the precision of the parameter estimate.

email: Valerii.V.Fedorov@gsk.com

HOW TO PREPARE THE BEST DATA PACKAGE FOR A DATA MONITORING COMMITTEE

Vipin Arora*, Novartis Pharmaceutucal Corporation
David Manner, Eli Lilly and Company

Preparing for a Data Monitoring Committee (DMC) can be very challenging from a Statistician's perspective. The data package presented to the DMC is the key to assess patient safety and also ensure that studies that are very likely to fail are stopped earlier (due to futility). Similarly, the studies that involve investigational drugs/devices that are very likely to fulfill huge unmet medical need(s) are stopped sooner to allow Sponsor to consider submissions of new drug/device(s) to the Health Authorities (stopped for efficacy). Often, the DMCs are faced with the challenge of too much information being submitted for their review. This information although overwhelming, some critical information and observed relationships between efficacy and safety parameters may still be missing. This gap could pose severe limitations to the DMC for an informative recommendation to the sponsor. Statisticians therefore have a huge responsibility to plan, implement and present an optimal data package to ensure efficient decision by the DMC as per the Charter under the study protocol. The data preparations tasks may span over long duration and may involve outside vendors (CROs) and adjudication committees due to regulatory requirements. A few examples of such data preparations will be included in this paper based on pharmaceutical industry experience. Other topics, e.g., communication flow between DMC and Sponsor and follow up actions, will also be presented.

email: vipin.arora@pharma.novartis.com

EVALUATION OF THE QUALITY OF INVESTIGATIVE CENTERS USING CLINICAL RATING AND COMPLIANCE DATA

Junyuan Wang* , Merck Research Laboratory
Junfeng Sun, The Ohio State University
Guanghan Liu, Merck Research Laboratory

In late phase clinical trials, a large number of investigative centers are usually needed in order to recruite sufficient number of patients in a timely manner. The quality of the studies performed by the investigative centers is crucial for the success of the trial. If we can identify the centers that perform poorly, and avoid using them in future trials, we may improve our chance of success. We explored the possibility of evaluating clinical centers based on combination of the clinically rated scores and patient compliance data. The data came from four double-blind, multi-center, placebo and active controlled studies of an investigational new drug for treatment of major depressive disorder. We derived variables for the compliance using either summary or model based estimates. These compliance variables include visit compliance, medication compliance, percent of protocol violator, and percent of dropouts. A new overall rating scheme of centers is proposed by combining the compliance variables and clinical rated scores using principle component analysis. This new rating method is compared with the clinical overall rating, and justifications for the new rating are provided.

email: junyuan_wang@merck.com

DETECTING QUALITATIVE INTERACTIONS IN CLINICAL TRIALS: AN EXTENSION OF RANGE TEST

Jianjun (David) Li*, Merck
Ivan Chan, Merck

When a conclusion on the treatment effect is to be made in clinical trials, it is natural to ask if the treatment effect is the same in various subsets of patients. The qualitative interaction, which means that the treatment is beneficial in some subsets and harmful in the others, is generally of major importance. In this paper, a new statistical test is developed for detecting such interactions. The new test is an extension of the well- known range test, but utilizes all observed treatment differences rather than only the maximum and minimum values. The extensive simulations concludes the proposed extended range test generally outperforms the range test, and is advantageous even to the likelihood ratio test, when the prior knowledge on distribution of treatment effect over the patient subsets is not available, as often is the case in practice. It is also illustrated through data from a clinical trial that the extended range test detects the qualitative interaction while the range test and likelihood ratio test do not.

email: jianjun_li@merck.com

COVARIANCE TAPERING FOR LIKELIHOOD-BASED ESTIMATION IN LARGE SPATIAL DATA SETS

Cari Kaufman*, Carnegie Mellon University
Mark Schervish, Carnegie Mellon University
Doug Nychka, National Center for Atmospheric Research
Reinhard Furrer, National Center for Atmospheric Research

Maximum likelihood and other likelihood based methods, such as REML or Bayesian methods, are attractive approaches to estimating covariance parameters in spatial models based on Gaussian random fields. Finding such estimates can be computationally infeasible for large datasets, however, requiring O(N^3) calculations for each evaluation of the likelihood based on N spatial locations. We use the method of covariance tapering to approximate the likelihood in this setting. The model and sample covariance are both directly multiplied by a compactly supported covariance, giving matrices which can be be manipulated using sparse matrix algorithms. We explore the efficiency of the resulting estimators and their value as plug-in estimates for kriging through numerical simulation and a study of their robust information criterion.

email: cgk@stat.cmu.edu

RECURSIVE PARTITIONING IN SPATIALLY CORRELATED DATA

Patrick S. Carmack*, University of Texas Southwestern Medical Center
William R. Schucany, Southern Methodist University

Recursive partitioning is a general technique for dividing data in homogeneous regions. The meaning of homogenous changes depending on the application. In the case of brain images, homogeneity can be defined in terms of spatial correlation. Using this technique, brain images can be partitioned into spatially homogenous regions with a localized spatial Kriging model in each region. Unfortunately, brain image packages like SPM fail to exploit spatial correlation in brain image data, but yield a comprehensive analysis of the entire brain. Spatial Kriging models have been fit in manually selected spatially homogenous regions of the brain. While these models take advantage of spatial correlation, they fail to yield a comprehensive analysis of the entire brain. Recursive partitioning promises to automate the placement of localized spatial Kriging models. These models exploit the correlation data and potentially yield an analysis of the entire brain.

email: patrick.carmack@utsouthwestern.edu

SPATIAL STOCHASTIC VOLATILITY

Huiliang Xie*, University of Iowa
Jun Yan, University of Iowa

A widely used spatial areal model (Besag, York, and Mollie, 1991, Ann. Inst. Stat. Math.) assumes that the error is composed of two terms: one capturing the regional clustering and the other representing regional heterogeneity. The regional heterogeneity terms are assumed to be independent and identically distributed normal variables, which may fail to deliver the spatial heteroskedasticity arising in many real datasets and, therefore, fail to provide good prediction intervals. This paper introduces a new class of spatial processes with spatial stochastic volatility to model the regional heterogeneity. These are mean zero, conditionally independent processes given a latent spatial process of the variances. The logarithm of the latent variance process can be modeled using a conditional or intrinsic conditional autoregression. The spatial stochastic volatility (SSV) model relaxes the traditional homeskedasticity assumption for spatial heterogeneity and brings great flexibility to the popular spatial statistical models. This model framework has an even larger number of random effects in a hierarchical setting than the existing spatial areal models, making the estimation of the parameters a challenging problem. An efficient structured Markov chain Monte Carlo algorithm is developed to update the parameters in blocks, using a normal mixture approximation of the distribution of a log-Chi-square variable. A simple nonparametric test is proposed to detect volatility clustering. The spatial stochastic volatility models can be applied to many applications where spatial heteroskedasticity may present. The well-known dataset of wheat yield is used to illustrate how spatial stochastic volatility can improve over the existing analysis.

email: huxie@stat.uiowa.edu

BAYESIAN AREAL WOMBLING FOR GEOGRAPHICAL BOUNDARY

Haolan Lu*, University of Minnesota
Brad Carlin, University of Minnesota

In the analysis of spatially referenced data, interest often focuses not on prediction of the spatially indexed variable itself, but on boundary analysis, i.e., the determination of boundaries on the map that separate areas of higher and lower values. Existing boundary analysis methods are generically referred to as wombling, after a foundational paper by Womble (1951). When data are available at point level, such boundaries are obtained by locating the points of steepest ascent or descent on the fitted spatial surface. In this paper we propose related methods for areal data. Such methods are valuable in determining boundaries for datasets that are available only in ecological format. After a brief review of existing algorithmic techniques, we propose a fully model-based framework for areal wombling, using Bayesian hierarchical models. We explore the suitability of various existing hierarchical and spatial software packages to the task, and show the approach's superiority over existing non-stochastic alternatives, both in terms of utility and average mean square error behavior. We also illustrate our methods using Minnesota colorectal cancer late detection data.

email: haolanl@biostat.umn.edu

SIMULTANEOUS CONFIDENCE INTERVALS FOR RATIOS OF NONPARAMETRIC INTENSITY

Traci l. Leong*, Rollins School of Public Health, Emory University
Lance A. Waller, Rollins School of Public Health, Emory University

The intensity function for a spatial point process defines the number of events expected per unit area. Ratios of intensity functions for two spatial point processes (e.g., cases and controls) over the same study area serve as local measures of relative risk. Current approaches for local interval estimation of the relative risk surface based on ratios of kernel intensity estimates provide pointwise inference. We explore approaches for simultaneous interval estimation of the relative risk surface incorporating spatial correlation between pointwise intervals induced by the kernel bandwidth. We illustrate the approaches using sea turtle nesting locations on Juno Beach, Florida.

email: tleong@sph.emory.edu

PROCESS CONVOLUTION APPROACH TO RECONSTRUCTION OF BINARY FIELDS

Margaret B. Short*, Los Alamos National Lab
Dave Higdon, Los Alamos National Lab

We discuss a process convolution approach to estimating binary fields, implemented via Markov chain Monte Carlo. An archeological data set is used to locate regions of human activity. A geological data set from an Italian aquitard is used to determine regions of high and low permeability. This represents on-going work with David Higdon, Daniel Tartakovsky and Alberto Guadagnini.

email: mbshort@lanl.gov

SPATIAL ESTIMATION WITH COMPUTER-EFFICIENT PARSIMONIOUS INTERACTION MODELS

Ernst Linder*, University of New Hampshire

We propose a parsimonious class of computer-efficient Gaussian spatial interaction models that includes as special cases CAR and SAR - like models. In this class the spatial structure is parameterized similarly as in the Matern class of geostatistics. Furthermore we show that, for rectangular lattices, this class is equivalent to higher-order Markov random fields that were proposed by Rue and Tjelmeland (2002) to approximate geostatistical models. Thus we capture the computational advantage of iterative updating of Markov random fields, while at the same time providing the possibility of simple interpretation of the spatial structure analogous to the Matern class of spatial covariance functions. For very large data on a rectangular lattice this class lends itself for circular embedding / spectral basis expansion. This class of spatial models is defined via a spatial structure removing orthogonal transformation. The latter is a one-time preprocessing step in iterative estimation, such as MCMC. We show how this model is embedded in spatial and spatial temporal applications. Keywords: Markov random fields, CAR models, spatio-temporal modeling

email: elinder@cisunix.unh.edu

---

98. SURVIVAL ANALYSIS II

APPLICATION AND ASSESSMENT OF RESIDUAL-BASED CLASSIFICATION APPROACHES ON MELANOMA SURVIVAL DATA: VALIDATION OF AJC ON CANCER MELANOMA STAGING SYSTEM

Chen-An Tsai*, UAB Comprehensive Cancer Center
Dung-Tsa Chen, UAB Comprehensive Cancer Center
Seng-Jaw Soong, UAB Comprehensive Cancer Center

Classical Cox proportional hazards model has been widely used in biomedical area to identify key prognostic factors in tumor patients, such as tumor thickness and lesion location. However, the analysis results based on the identified prognostic factors seem not so practical for clinical investigators because it does not provide explicit criteria to classify patients into different risk groups. In this study, we apply two well-known classification schemes in microarray data analysis, regression tree and random forest, to develop clinical prediction rules on melanoma survival data. We will demonstrate these two approaches are suitable to more general situations (e.g., grouping homogeneous patients and predicting a new patient population) than classical Cox models. In order to apply classification approaches to right censored survival data, we will use null martingale residuals (Therneau et al., 1990) or deviance residuals from Cox proportional hazards model as the response variables. In addition, two measures, Integrated Brier score and Explained residual variation, are used to compare classification schemes. Analysis results from the two approaches will be used to validate the cancer melanoma staging system.

email: ctsai@uab.edu

LOCAL LINEAR ESTIMATION OF A SMOOTH DISTRIBUTION BASED ON CENSORED DATA

Liang Peng, Georgia Institute of Technology
Shan Sun*, Texas Tech University

We propose a local linear estimator of a smooth distribution function based on censored data. This new estimator applies local linear techniques to observations from a regression model where the value of the product limit estimator equals the value of the true distribution plus an error term. We show that for most commonly used kernel functions, our local linear estimator has a smaller mean squared error than the kernel estimator proposed by Ghorai and Susarla (1990).

email: ssun@math.ttu.edu

MULTIPLE AUGMENTATION WITH OUTCOME DEPENDENT SAMPLING

Shuangge Ma*, University of Washington

Outcome dependent sampling schemes have been proposed in large epidemiologic cohort studies as a way of enlarging the relative size of subsamples of interest and reducing costs associated with assembly of covariate histories, especially when the events of interest are rare. Well known examples of outcome dependent sampling schemes include the case-cohort sampling proposed in Prentice (1986), case control sampling (Breslow and Day, 1980) and the randomized recruitment sampling investigated in Weinberg and Sandler (1991). In this paper, we propose a general semiparametric multiple data augmentation approach for outcome dependent sampled data, when the outcome is known for the whole cohort. Computational algorithms for the Poor Man's and the Asymptotic Normal data augmentations are investigated. Simulation studies show that the data augmentation approach is superior compared with existing approaches, while still being computationally affordable. We apply the proposed technique to the analysis of the South Wales Nickel Worker Study data, from which we simulate case-cohort datasets, and the Multi-Ethnic Study of Atherosclerosis (MESA) data, where the covariates are known on a randomized recruitment basis.

email: shuangge@biostat.wisc.edu

SURVIVAL MODEL AND ESTIMATION FOR LUNG CANCER PATIENTS

Xingchen A. Yuan*, East Tennessee State University
Don Hong, Vanderbilt University
Shyr Yu, Vanderbilt University

Lung cancer is the most frequent fatal cancer in the United States. By assuming a form for the hazard function for a group of lung cancer patients, the covariates in the hazard function are estimated by maximum likelihood estimation following the proportional hazards regression analysis. Although the proportional hazards model does not give the explicit baseline hazard function, it can be determined by fitting the data with a non-linear least square technique. The survival model is then examined by a neural network simulation. The neural network learns the survival pattern from available hospital data and gives survival prediction for random covariate combinations. The simulation results support the covariate estimation in the survival model.

email: xingchenyuan@hotmail.com

MODELING RELIGION'S INFLUENCE ON HIV PROGRESSION AND MORTALITY

John A. Myers*, Yale School of Medicine
Musie Ghbremicheal, Yale School of Medicine
Heping Zhang, Yale School of Medicine

Religion may have an impact on health and disease progression. Religion can bring a sense of meaning and purpose to individuals. Evidence suggests that religion and spirituality is an effective way to cope with the psychological impact of HIV infection. However, empirical evidence demonstrating the influence of religion on the physical health status of those infected with HIV is scarce. While there is evidence that religious behavior (e.g. service attendance, prayer, reading religious literature) is associated with higher CD4+ counts in gay men, there have been no studies focused on the association of religion on viral load and mortality rates. Furthermore, no such analyses have been conducted in women. Women use religion as a coping strategy to deal with illness more so than men. We used the HIV Epidemiology Research Study (HERS) data set, comprised of 2813 HIV positive women and 989 HIV negative women, to examine the relationship between religion and viral load, as well as mortality rates, in women infected with HIV. We model religion's influence on CD4+ count, viral load, and mortality rates in women infected with HIV using parametric and non-parametric regression techniques.

email: john.myers@yale.edu

ON A CONNECTION BETWEEN THE PARAMETRIC LIKELIHOOD AND THE EMPIRICAL LIKELIHOOD

Min Chen*, University of Kentucky
Mai Zhou, University of Kentucky

Maximum likelihood method is a widely used statistical inference method. The maximum parametric likelihood estimator is well known for its nice properties such as efficiency and consistency. The recently developed empirical likelihood method, (Thomas and Grunkmier 1975, Owen 2001), is a nonparametric procedure. Empirical likelihood and parametric likelihood are closely related, so it is natural to expect a close relationship between the maximum empirical likelihood estimator and the parametric maximum likelihood estimator. We illustrate the close relationship between the maximum empirical likelihood estimator and the parametric maximum likelihood estimator by examine a right censored data estimation problem. We show that the variance of the maximum empirical likelihood estimator and the variance of the parametric maximum likelihood estimator are similar. We also show that correlation between the two estimators is positive and can be made arbitrarily close to 1 if proper restrictions are imposed on the empirical likelihood. This is similar but in opposite order to the sieve method.

email: minchen@ms.uky.edu

ONE- AND TWO-SAMPLE NONPARAMETRIC INFERENCE PROCEDURES IN THE PRESENCE OF DEPENDENT CENSORING

Yuhyun Park*, Harvard University
Lu Tian, Northwestern University
L. J. Wei, Harvard University

In survival analysis, the event time T is often subject to dependent censorship. Without assuming a parametric model between the failure and censoring times, the parameter Θ of interest, for example, the survival function of $T,$ is generally not identifiable. On the other hand, the collection Ω of all attainable values for Θ may be well-defined. In this article, we present non-parametric inference procedures for Ω in the presence of a mixture of dependent and independent censoring variables. By varying the criteria of classifying censoring to the dependent or independent category, our proposals can be quite useful for the so-called sensitivity analysis of censored failure times. The case that the failure time is subject to dependent interval censorship is also discussed in this article. The new proposals are illustrated with data from two clinical studies on HIV-related diseases.

email: ypark@hsph.harvard.edu

99. ASSOCIATIVE ANALYSIS OF GENETIC DATA

INCORPORATING CLUSTERING UNCERTAINTY IN REGRESSION-BASED ANALYSIS FOR HAPLOTYPE-DISEASE ASSOCIATION

Jung-Ying Tzeng*, North Carolina State University

Genetic variants of complex diseases are governed by multiple potentially interacting genetic and environmental factors. To map the genes underlying complex diseases, regression-based association methods are becoming increasingly important in modern haplotype analyses. Such methods can model various types of clinical phenotypes; they can also accommodate environmental covariates, polygenic effects and the interactions among them. By treating haplotypes as covariates, the regression-based methods can evaluate the etiological effects of individual haplotypes. However, the practical efficacy of model-based approaches is limited by potentially large number of parameters for modeling haplotype diversity. To increase the power and efficiency of haplotype analysis, we have introduced an evolutionary-based clustering algorithm to group haplotypes (Tzeng 2004; Genet Epidemiol). This clustering algorithm takes the uncertainty of underlying evolutionary relationship into account and assigns each haplotype a probability vector for the corresponding clusters. Here we describe its extension to regression approaches for assessing the haplotype-disease association. Our generalized linear model incorporates the clustering algorithm into the framework. We will discuss how to infer the effect of clusters of homogeneous haplotypes and evaluate their significance.

email: jytzeng@stat.ncsu.edu

USING TREE-BASED RECURSIVE PARTITIONING METHODS TO GROUP HAPLOTYPES IN ASSOCIATION STUDIES

Kai Yu*, Washington University, St. Louis
Jun Xu, Procter & Gamble Co.
D. C. Rao, Washington University, St. Louis
Michael Province, Washington University, St. Louis

Motivated by the increasing availability of high-density single nucleotide polymorphisms markers across the genome, various haplotype-based methods have been developed for candidate gene association studies, and even for genome-wide association studies. A negative feature of such haplotype-based methods is the relatively large number of existing haplotypes, which increases the degrees of freedom and decreases the power for the corresponding test statistic. To limit the degrees of freedom, we propose a procedure that uses a tree-based recursive partitioning algorithm to group haplotypes into a small number of clusters, and conducts the association test based on groups of haplotypes, instead of individual haplotypes. The method can be used for both population-based and family-based association studies, with known or ambiguous phase information. Simulation studies suggest that the proposed method has the right type I error rate, and is more powerful than some existing haplotype- based tests

email: kai@wubios.wustl.edu

INFERENCE UNDER VARIOUS MULTINOMIAL DISTRIBUTION MODELS FOR HAPLOTYPES UNDER HARDY-WEINBERG DISEQUILIBRIUM

Beverly M. Snively*, Wake Forest University School of Medicine
David M. Reboussin, Wake Forest University School of Medicine
Christine E. McLaren, University of California, Irvine
Ronald T. Acton, University of Alabama at Birmingham
Mark R. Speechley, University of Western Ontario
Emily L. Harris, Kaiser Permanente Northwest
James C. Barton, Southern Iron Disorders Center
Cathie Leiendecker-Foster, University of Minnesota
Victor R. Gordeuk, Howard University

Multinomial probability distribution models provide for statistical inference on population genotype frequencies. When Hardy-Weinberg (HW) disequilibrium is observed, genotype frequencies are generally estimated from genotypic counts using multinomial likelihood estimators. If the disequilibrium derives from selection bias, simplicity of the usual multinomial model could be outweighed by improved properties of alternate models. The goals of this study are to 1) describe alternate models for genotype or haplotype (or allele) data that allow for selection, 2) develop hypothesis tests and estimation procedures for making inferences about genotype frequencies, and 3) compare the usual and alternate models, applied to HFE C282Y and H63D genotypes from the Hemochromatosis and Iron Overload Screening (HEIRS) Study. Four models were studied: two usual and one conditional multinomial model, and a mixture model with multinomial components. Score tests and corresponding confidence intervals were calculated. C282Y genotypes deviated from HW equilibrium in subgroups of whites and Hispanics in the HEIRS Study–possibly due to greater participation among eligible C282Y homozygotes. Conditional and mixture models yielded lower frequency estimates for C282Y homozygotes in these subgroups. Further investigation is needed on conditions under which these alternate models are preferred.

email: bmellen@wfubmc.edu

INFERENCE ON HAPLOTYPE-ENVIRONMENT INTERACTIONS USING GENOTYPE DATA FROM CASE-CONTROL STUDIES

Lydia C. Kwee*, Emory University
Amita K. Manatunga, Emory University
Glen A. Satten, Centers for Disease Control and Prevention
Michael P. Epstein, Emory University

Genetic association studies often use a case-control study design due to the ease of sample collection. Statistical methods that utilize prospective likelihoods for analyzing retrospective genetic data may suffer from a loss of statistical efficiency in certain situations, so it is of interest to use retrospective likelihoods when feasible. For haplotype analysis of case-control genotype data, Epstein & Satten (2003) proposed a retrospective likelihood approach that allowed for testing of both global and specific haplotype effects on disease. In Satten and Epstein (2004), they showed that their retrospective approach had optimal power for detecting haplotype-disease association relative to related prospective approaches. In the current work, we extend this powerful retrospective approach to allow for covariates, which permits us to model and test main environmental effects as well as gene-environment interaction effects. In order to accommodate the existence of ambiguous haplotypes in the genotype data, we apply a variant of the EM algorithm for proper inference. The power of this modified approach is compared to that of the prospective haplotype approach of Schaid et al. (2002). The new method is also applied to case-control data from the Finland-United States Investigation of Non-Insulin Dependent Diabetes Mellitus (FUSION) genetic study.

email: lkwee@sph.emory.edu

JOINT LINKAGE AND ASSOCIATION MAPPING OF QUANTITATIVE TRAIT LOCI, A HAPLOTYPE-BASED APPROACH

Ruzong Fan*, Texas A&M University
Jeesun Jung; University of Pittsburgh
Lei Jin, Texas A&M University

Variance component models are proposed for high resolution joint linkage and association mapping of quantitative trait loci (QTL), based on both pedigree and population haplotype data. Suppose that a quantitative trait locus is located in a chromosome region. In the region, a haplotype block is typed which may consist of several markers such as single nucleotide polymorphisms (SNPs). Suppose that a sample is available which consists of both pedigree and population data. Two regression models, ``genotype effect model'' and ``additive effect model'', are proposed to model the association between the haplotype block and the trait locus. The linkage information, i.e., recombination fraction between the QTL and the haplotype block, is modeled in the variance and covariance matrix. By analytical formulae, we show that the ``genotype effect model'' can be used to model the additive and dominant effects simultaneously; the ``additive effect model'' only takes care of additive effect. Based on the two models, F- test statistics are proposed to test association between the QTL and haplotype block. The non-centrality parameter approximations of F-test statistics are derived to make power calculation and comparison.

email: rfan@stat.tamu.edu

QUANTIFYING BIAS DUE TO GENOTYPING ERROR IN CASE-CONTROL STUDIES OF GENE HAPLOTYPES AND CANCER

Usha S. Govindarajulu*, Harvard University
Donna Spiegelman, Harvard University
David J. Hunter, Harvard University
Peter Kraft, Harvard University

Genotyping errors can induce small biases in haplotype frequency estimates. Here we consider the impact of genotyping error on haplotype odds ratio estimates from case-control studies of unrelated individuals. We calculate asymptotic bias analytically, deriving the exact likelihood of the observed data given a model for genotype misclassification. For simplicity, we assume phase is known and that genotyping errors occur independently and at the same rate at each locus. We use empirical haplotype frequencies from SeattleSNPs (http://pga.gs.washington.edu/) for genes with diverse structure (one to eight haplotype blocks). We find that for common haplotypes (> 5% frequency), realistic genotyping error rates (<= 1%), and moderate relative risks (e.g. 2) the asymptotic bias due to genotyping error is small (<5%) and directed towards the null. The magnitude of the bias is a function of haplotype frequency (the smaller the frequency, the greater the bias) as well as the similarity among common haplotypes. OR's for less common haps (5% and less) are more strongly affected. Furthermore, genotyping error increased the cumulative frequency of very rare haplotypes (most of which were spurious). This suggests multi-locus genotypes may be a useful quality control tool, as very rare haplotypes may be the result of genotyping errors.

email: usha@alum.bu.edu

ACCOUNTING FOR POPULATION STRATIFICATION IN CASE-CONTROL STUDIES OF GENETIC ASSOCIATION: A BAYESIAN APPROACH

Li Zhang*, University of Florida
Bhramar Mukherjee, University of Florida
Malay Ghosh, University of Florida
Rongling Wu, University of Florida

We propose a parametric Bayesian approach to examine the association between a candidate gene and the occurrence of a disease after accounting for population stratification. We account for population substructure by using information on additional marker loci whose alleles vary in frequency among subpopulations. Our approach does not depend on the assumption of linkage equilibrium which was required in previous studies. In particular, we extend our model to address the situation (i) when one or more of the loci could be in linkage disequilibrium with the candidate gene, (ii)when the disease may not be rare, the candidate gene may not be bi-allelic and can have any number of genotypes. The proposed Bayesian procedure is implemented via Markov chain Monte carlo numerical integration technique. Estimation results based on a simulated admixed population (mimicking the results presented in Sala et al. (1998,1999)) shows that the estimates of the relative risk parameters using additional mutilocus genetic information are superior to those when one does not exploit such information in the model.

email: mlizhang@ufl.edu

NOTES

ENAR WEBSITE
WWW.ENAR.ORG

POSTER SESSION

SOFTWARE FOR CHOOSING SAMPLE SIZE FOR CONFIDENCE INTERVAL-BASED INFERENCES IN GAUSSIAN LINEAR MODELS

Michael R. Jiroutek*, Bristol-Myers Squibb Pharmaceutical Research Institute
Keith E. Muller, University of North Carolina at Chapel Hill

We describe free software for choosing a sample size for confidence interval based inferences for any scalar parameter in General Linear Multivariate Models. The software implements the methods of Jiroutek, Muller, Kupper and Stewart (2003). Historically, the sample size chosen for a study aimed at estimating a confidence interval controlled only the probability of width (the confidence interval is as narrow as desired). In contrast, the criterion introduced by Jiroutek, et al. (2003) controls the probability of width and rejection (of the null hypothesis), given validity (the interval contains the true, unknown parameter). Many previously published power and confidence interval criteria are special cases. The new approach better aligns the sample size rule with common scientific goals. Numerical examples computed with the new program illustrate the dramatic errors that can be made when sample size computations are not properly aligned with study outcomes. All combinations of one- and two-sided confidence intervals and hypothesis tests are handled appropriately. The program is freely available at http://www.bios.unc.edu/~muller.

email: michael.jiroutek@bms.com

GENERALIZED MULTI-LOCUS SCORE STATISTICS FOR THE CASE-CONTROL ASSOCIATION STUDIES

Qing Lu*, Case Western Reserve University
Tao Wang, Case Western Reserve University
Chao Xing, Case Western Reserve University
ZhiYing Xu, Case Western Reserve University
Katrina A. Goddard, Case Western Reserve University

With the increasing number of SNPs and the higher density of markers that are available, association studies have growing potential for dissecting the genetics of common disease. Meanwhile, the increasing density of markers creates new challenges in association studies, including inter-marker correlation, multiple testing, and interactions between SNPs. Here we propose a generalized linear model based multi-locus score statistics method for the joint analysis of multiple correlated marker loci. By incorporating the flexible moving window technique, best subset selection, permutation and gene-gene interaction, our model is able to handle these issues. In addition, this method offers flexibility by including either dichotomous or quantitative traits, and environmental or gene-environment interaction variables. In a simulation study, we compare our model with the sum statistics [Hoh J et al 2001 Genome Res] for a variety of linkage disequilibrium and disease model settings. In additive model scenario with high linkage disequilibrium, when the type I error is controlled at 0.01 level, the power of our model is 0.792 while that of sum statistics is 0.776. Our method deals with the anticipated complexities for association analysis with large numbers of markers, and will limit the false positive rate while maintaining power to detect disease loci.

email: qlu@darwin.cwru.edu

IMPROVED GROUP TESTING ESTIMATION OF TRAIT PREVALENCE

Christopher R. Bilder*, University of Nebraska-Lincoln
Joshua M. Tebbs, Kansas State University

Group testing is used to estimate trait prevalence in situations where individual testing is too costly or logistically unfeasible. Estimating the prevalence via maximum likelihood has been the usual approach, but unfortunately the resulting estimator is positively biased. We propose new parametric empirical Bayesian estimators that reduce this bias. Furthermore, these estimators are also shown to usually have smaller mean squared error and Bayes risk than the maximum likelihood estimator. We illustrate these new estimators using data from a multiple-vector-transfer experiment involving the Mal Rio Cuarto virus, where the goal is to estimate the probability of transmission by the Delphacodes kuscheli planthopper.

email: cbilder3@unl.edu

ISSUES WITH USING THE KAPPA STATISTIC TO COMPARE DIAGNOSTIC TESTS WITHOUT A GOLD STANDARD

Stephanie T. Broyles*, Louisiana State University Health Sciences Center
Leann Myers, Tulane School of Public Health and Tropical Medicine

The kappa statistic (k) has been used to describe the agreement between diagnostic tests in the absence of a gold standard, recently in providing guidelines for use of an alternative test for detecting latent tuberculosis infection (LTBI). The current research uses simulations to describe how kappa values are affected by disease prevalence, the two tests' sensitivities and specificities, the equivalence of the sensitivities and specificities of the two diagnostic tests, and the level of dependence of the diagnostic tests as a result of measuring the same underlying biological marker. Simulations also describe the sensitivity of the kappa statistic to being combined across populations. Because kappa values are related to many factors, values of this statistic are difficult to apply without without prior knowledge of the performance of the two diagnostic tests. For example, in the case of the comparison of the two tests for LTBI, guidelines may erroneously be recommending use of the inferior test. The kappa statistic is, therefore, not recommended for comparison of diagnostic tests in the absence of a gold standard, and researchers are advised to use methods that incorporate existing knowledge about the performance of the tests into a model that permits the estimation and comparison of the two tests' true diagnostic abilities.

email: sbroyl@lsuhsc.edu

HYPOTHESIS TESTING WITH TWO-PART MODELS

Leann Myers*, Tulane University
Yeonjoo Yi, Tulane University
Hao He, Tulane University

Previous analyses showed the utility of two-part models for prediction with epidemiological data (Myers & Yi, 2003 & 2004; Yi & Myers, 2004). When a single observation per subject is available, the strategy is to use logistic regression to predict who has a non-zero response (Part 1) and a linear regression to predict level of response among those with non-zero values (Part 2). The two parts are then combined to predict level of response for the entire sample. When there are multiple observations per subject, generalized estimating equations (GEE) with a logit link is used in Part 1 and GEE with a normal link is used in Part 2. These approaches were previously assessed in terms of prediction. The current study focused on hypothesis testing. Scenarios included covariates related to whether the response was non-zero, covariates related to level of response among non-zero data, and combinations of both. Bootstrapping methods were used to assess standard errors. Type I error rates and power were assessed for different sample sizes, different levels of non-response, and both the single and multiple response cases. The two-part model can be used for hypothesis testing as well as for prediction, although there are restrictions.

email: myersl@tulane.edu

AN EMPIRICAL COMPARISON OF THE POWER AND ACCURACY OF THREE SPECIFIC CLUSTER TESTS UNDER MISCLASSIFICATION ERROR

Leslie H. Morgan*, Tulane University
Leann Myers, Tulane University
Frances J. Mather, Tulane University

For cases that have been geocoded to residential address and aggregated to larger geographic areas, the counts are not completely accurate. An empirical study of the power and accuracy of three specific cluster detection methods was conducted under varying levels of misclassification error. The cluster detection tests were a simple outlier detection method based on hierarchical generalized linear model results from the GLIMMIX macro in SAS 9.1, Tango's maximized excess events test (MEET), and the spatial scan statistic in SaTScanTM. The population structure of the Louisiana counties was used to generate cases that follow a Poisson process with three fixed effects, age, year of diagnosis, and a county-level deprivation score. A random intercept model with a normally distributed county-level effect was simulated with three levels of variance. The first level was no variance attributable to the group level. The second and third levels of variance resulted in expected risk ratios between counties of 2.0 and 4.0, respectively. The levels of misclassification error were set at 0, 5, and 10 percent. The models were simulated with and without an urban cluster effect. For the 18 resulting models, the power, type I error rate, and accuracy of specific cluster location were examined.

email: lmorgan1@tulane.edu

LEARNING CURVES IN CLASSIFICATION WITH MICROARRAY DATA

Kenneth R. Hess*, University of Texas M. D. Anderson Cancer Center
David L. Gold, University of Texas M. D. Anderson Cancer Center

Learning curves are used to model improvements in task performance with accumulated experience. This concept has been adapted to characterize improvements in classifier performance with increasing numbers of training samples. Recently, learning curves fit with power-law models have been used effectively with microarray data. We fit learning curves to performance data from 12 published microarray datasets using a variety of classifiers. Our results confirm that the power law model captures trends in classification test performance given the training sample size, and that a reasonable prediction of performance is achievable across a variety of microarray datasets and binary classifiers with at least 65 samples. We find that: (1) learning curves sometimes cross, indicating that the preferred classifier may depend on sample numbers available for training, (2) outcome studies are generally harder to classify than other kinds of microarray data, and (3) bootstrap resampling can yield approximate prediction intervals. Preliminary results show that learning curves can also be used for conventional prognostic factor studies using for example the Cox proportional hazards regression model to characterize how prediction accuracy changes with sample size. Learning curves are useful for estimating the number of training samples needed to achieve a given level of accuracy.

email: khess@mdanderson.org

STANDARD ERRORS FOR ATTRIBUTABLE RISK FOR SIMPLE AND COMPLEX SAMPLES

Barry I. Graubard*, National Cancer Institute
Thomas R. Fears, National Cancer Institute

Adjusted attributable risk (AR) is the proportion of diseased individuals in a population that is due to an exposure. We consider estimates of adjusted AR based on odds ratios from logistic regression to adjust for confounding. Influence function methods used in survey sampling are applied to obtain simple and easily programmable expressions for estimating the variance of $\widehat{AR}$. These variance estimators can be applied to data from case-control, cross-sectional and cohort studies with or without frequency or individual matching and for sample designs with subject samples that range from simple random samples to (sample) weighted multi-stage stratified cluster samples like those used in national household surveys. The variance estimation of $\widehat{AR}$ is illustrated with: (i) a weighted stratified multistage clustered cross-sectional study of childhood asthma from the Third National Health and Examination Survey, and (ii) a frequency-matched case-control study of melanoma skin cancer.

email: graubarb@mail.nih.gov

A COMPARISON OF ASSOCIATION STUDIES FOR HAPLOTYPE RISK FACTORS USING PHASE UNKNOWN GENOTYPE DATA

Jingxia Liu*, Medical College of Wisconsin
Tao Wang, Medical College of Wisconsin

In association mapping studies, there is an increased interest in haplotypes as opposed to single markers due to the fact that a gene's protein product may more likely be induced by numerous mutations over a chromosomal region. Two different methods have been introduced in the analysis of the haplotype-phenotype relationships. One available method is to estimate the haplotype frequencies from marker genotype data first using either the expectation-maximization algorithm or genotype information from relatives. Then apply a weighted regression on the haplotypes to identify association of the haplotypes with a phenotype. Another method is to estimate the haplotype frequencies and association effects of these haplotypes jointly. In this article, we compared the two methods and showed that there are some disadvantages in some cases. We also developed a latent variable approach for association analysis of haplotypes. An EM algorithm was established to distinguish phases using phase-unknown genotype data. The latent variable model can not only capture the risk haplotypes but also provide us a general way of testing for association between phenotype and marker-genotypes. In addition, a comparison of the three methods was made through extensive simulation studies.

email: liujingxia@yahoo.com

BAYESIAN HIERARCHICAL MODELING IN DIABETES QUALITY OF CARE STUDIES

Theodore J. Thompson*, Centers for Disease Control and Prevention
James P. Boyle, Centers for Disease Control and Prevention

Objectives: To model LDL cholesterol levels as a function of quality of care, and to estimate the relationship of cholesterol level with a cluster level quality of care score in persons with diabetes. Methods: Bayesian hierarchical regression models including random effects for cluster and person were fit using Markov chain Monte Carlo methods (MCMC). Non- informative prior distributions were used for the fixed effects and variance components. Markov chain convergence was assessed via graphical and numeric diagnostics. Maximum likelihood estimates were used as starting values for the chain. Estimates of marginal means were used instead of regression model parameters to simplify interpretation of results. The data are from the Translating Research into Action for Diabetes (TRIAD) study. A prospective cohort of 7349 people from 71 clusters was measured at two time points (March 2001 and July 2002). There are 6557 observations at time 1 and 4350 observations at time 2. Conclusions: Although there is an association between quality of care and cholesterol at time 1, there is no association at time 2. Cholesterol levels are declining with time in this cohort. Bayesian hierarchical models are useful for analyzing these clustered longitudinal data. It is straight forward to calculate posterior distributions for any estimand of interest from the Markov chain simulations.

email: tat5@cdc.gov

POSTERIOR PREDICTIVE CHECKING OF BAYESIAN HIERARCHICAL MODELS: A CASE STUDY IN DIABETES QUALITY OF CARE

James P. Boyle*, Centers for Disease Control and Prevention
Theodore J. Thompson, Centers for Disease Control and Prevention

Bayesian hierarchical regression models were used to model LDL cholesterol as a function of quality of care in persons with diabetes. Posterior predictive checks were developed to test model consistency with the data. The entire data set is replicated for each posterior simulation. A test quantity is defined and its posterior distribution is compared to its observed value. Posterior predictive p- values are calculated. The test quantity should reflect aspects of the model relevant to the scientific issues being addressed. The test quantities used were mean LDL by cluster and time and proportion LDL > 130 mg/dl by cluster and time. Transformations of LDL values and alternative variance structures were considered for model improvement. Based on the test quantity mean LDL, all models considered performed well. A log transformation of LDL was required for an adequate fit with respect to the test quantity proportion LDL > 130. All statistical analyses should include checking the fit of the model to the data. In a Bayesian analysis using simulation, this is relatively easy. The hierarchical regression models fit well with respect to mean LDL values but needed improvement to fit proportion LDL > 130.

email: jboyle@cdc.gov

ESTIMATING VARIANCE AND CONFIDENCE INTERVAL FOR DISEASE PREVALENCE BASED ON DATA FROM POPULATION-BASED REGISTRIES

Limin X. Clegg*, National Cancer Institute, National Institutes of Health
Mitchell H. Gail, National Cancer Institute, National Institutes of Health
Eric J. Feuer, National Cancer Institute, National Institutes of Health

We propose a new ‘Poisson’ method to estimate the variance for prevalence estimates obtained by the counting method described by Gail et al.(1999) and to construct a confidence interval for the prevalence. We evaluate both the Poisson procedure and the procedure based on the bootstrap proposed by Gail et al. in simulated samples generated by resampling real data. These studies show that both variance estimators usually perform well and yield coverages of confidence intervals at nominal levels. When the number of disease survivors is very small, however, confidence intervals based on Poisson method have supra- nominal coverage, whereas those based on the procedure of Gail et al. tend to have below nominal coverage. For these reasons we recommend the Poisson method, which also reduces the computational burden considerably.

email: lin_clegg@nih.gov

EMPIRICAL BAYES METHOD FOR INCORPORATING DATA FROM MULTIPLE GENOME SCAN

T. Mark Beasley*, University of Alabama at Birmingham
Kui Zhang, University of Alabama at Birmingham
Howard Wiener, University of Alabama at Birmingham
Christopher I. Amos, University of Texas M. D. Anderson Cancer Center
David B. Allison, University of Alabama at Birmingham

Individual genome scans tend to have low power and can produce markedly biased estimates of QTL effects. Further, the confidence interval for their location is often prohibitively large for subsequent fine mapping and positional cloning. Given that a large number of genome scans have been conducted, not to mention the large number of variables and subsets tested, it is difficult to confidently rule out Type 1 Error as an explanation for significant effects even when there is apparent replication in a separate data set. We adapted Empirical Bayes (EB) methods (Morris 1983) to analyze multiple datasets simultaneously to alleviate each of these problems. We develop an EB meta-analysis method to integrate the same linkage statistic from multiple sib pair studies while not forcing all studies to have an identical marker map or a common estimated QTL effect. The updated linkage statistic then can be used for the estimation of QTL location and effect. Simulation results indicate that the EB method can account for the between-study heterogeneity and estimate the QTL location and effect more precisely as well as supply narrower confidence intervals than an individual study.

email: mbeasley@uab.edu

FREE POWER SOFTWARE FOR REPEATED MEASURES, MANOVA, AND SOME MIXED LINEAR MODELS USING SAS/IML (R)

Matthew J. Gribbin*, University of North Carolina at Chapel Hill
Jacqueline L. Johnson, University of North Carolina at Chapel Hill
Sean L. Simpson, University of North Carolina at Chapel Hill
Keith E. Muller, University of North Carolina at Chapel Hill

We describe the latest version of free software which provides convenient power calculations for a wide range of multivariate linear models with Gaussian errors. The ‘multivariate’ and ‘univariate’ approaches to repeated measures, as well as MANOVA tests, are covered. Power for a limited but useful range of mixed models may also be computed, with a careful and appropriate choice of inputs. F approximations are used throughout, and are reduced to exact forms whenever possible. The multivariate tests, Wilks, Pillai-Bartlett and Hotelling-Lawley, are covered as are univariate tests, Geisser-Greenhouse, Huynh-Feldt, Box conservative and uncorrected. Confidence limits may be requested for most power values to reflect the uncertainty due to using any combination of estimated variances and means. Errors found in previous software have been corrected. Automatically available output files simplify producing plots and tables for manuscripts. The documentation has been expanded to include a wider range of examples.

email: mgribbin@bios.unc.edu

REPEATED MEASURES FOR GAUSSIAN MULTIVARIATE LINEAR MODELS: A TUTORIAL

Sean L. Simpson*, University of North Carolina at Chapel Hill
Keith E. Muller, University of North Carolina at Chapel Hill
Chris S. Coffey, University of Alabama at Birmingham

Reasonably accurate power approximations for repeated measures and multivariate linear models with Gaussian errors have been available for over 15 years. However, without access to convenient software, study planners are often forced to settle for simple and inaccurate approximations. Limited training in task provides another important barrier. We seek to help by providing a detailed description of appropriate power analysis for a model with a two-way factorial model between subjects and one within subject factor. Power analysis for many popular repeated measures and multivariate designs, such as a two-group trial, can be treated as a special case of the one reported. We focus primarily on using a slightly updated version of the free power software POWERLIB in SAS/IML(r). The same designs are also evaluated with PASS(r) and nQUERY(r). Both "multivariate" and "univariate" approaches to repeated measures are covered. We emphasize the value of power curves for varying sample sizes and mean differences. Our results illustrate that although tests of polynomial trend are equivalent to tests of mean differences for the "multivariate" approach, different results can occur with the "univariate" approach.

email: simpsons@email.unc.edu

LIKELIHOOD-BASED EVALUATION OF NORMALIZATION METHODS

Bonnie J. LaFleur*, Vanderbilt University
Dean Billheimer, Vanderbilt University
Heidi Chen, Vanderbilt University

The need for normalization has been recognized in many areas of science, and competing methods have already been incorporated into software used to evaluate samples from microarray experiments. Methods for mass spectroscopy are not as well established and are currently being developed. Despite this work the field lacks an analytical and computational framework to evaluate proposed normalization methods. We have generalized a framework grounded in statistical principles. The analytic framework for normalization that we propose can be viewed as extension of widely used statistical techniques and is based on the Box-Cox method of transformations that allows us to express current methods of normalization in a common mathematical structure along with a quantitative measure of performance. Examples will be presented with computational evaluation of normalization techniques. We will include some popular methods being used in Raman spectroscopy, but the techniques can be used more generally and can include normalization techniques currently applied to MALDI-TOF mass spectroscopy, microarray analyses, radioactive labeling experiments and blotting techniques.

email: bonnie.lafleur@vanderbilt.edu

PERFORMANCE OF A LONGITUDINAL TWO-PART MODEL WHEN DATA ARE NONIGNORABLY MISSING

Yeonjoo Yi*, Tulane University
Leann Myers, Tulane University

Semicontinuous variables have a proportion of responses which are a single value (often zero) while the remaining responses follow a continuous, often skewed, distribution. In a two-part model, the first level models the probability that the semicontinuous variable takes on its point mass value, and the second level models the distribution of the variable given that it is not at its point mass. The two parts are then combined into a single model. In a previous series of analyses (Yi & Myers, 2004), GEE methods were used to extend the two-part model to longitudinal data characterized by ignorably missing responses. In these analyses, the two-part GEE models performed well in terms of accuracy of prediction over all conditions. In the current study, the GEE two-part model approach was extended to longitudinal data characterized by non-ignorable missing responses in both the logit and linear stages. Simulations were used to investigate the performance of the two-part model, which was assessed in terms of deviations between observed and predicted responses.

email: yyi@tulane.edu

GROWTH MODELS FOR MULTILEVEL MULTIVARIATE ORDINAL DATA

Eisuke Segawa*, University of Illinois at Chicago

Bayesian MCMC is applied to high-(8 to 15) dimensional hierarchical generalized linear models. It is known that commonly used methods such as penalized quasi-likelihood and marginal quasi-likelihood produce biased results if the conditional joint densities of the responses belonging to the clusters are away from normal. Although Gauss- Hermite quadrature (including adaptive one) works often better, it is computationally prohibitive for high- dimensional problems (e.g., more than 6 dimensions). We demonstrate that MCMC works well with reasonable computational time through real and simulated data. Growth models for multilevel multivariate ordinal data in a prevention study are used in our presentation.

email: esegawa@uic.edu

SOFTWARE FOR ANALYZING EFFICIENT COHORT SUBSAMPLING DESIGNS: RELATIVE, ABSOLUTE, AND ATTRIBUTABLE RISKS

Hormuzd A. Katki*, National Cancer Institute, NIH
Steven D. Mark, National Cancer Institute, NIH

In efficient subsampling of cohort studies, the censored lifetime outcomes and easily obtainable covariates are observed on all members, but certain expensive or difficult to obtain covariates are observed only on a subsample. Examples include two-phase, two-stage, case-cohort, nested case-control, and double sampling designs. Using Robins, Rotnitzky, and Zhao (JASA 1994), Mark (JSM Proceedings 2003, 2675-91) and Mark and Katki (JASA, in press) we present classes of estimators of relative, absolute, and attributable risks from Cox models and Kaplan-Meier curves. The validity of the estimators depends on correct specification of a model for the sampling probabilities of the sampled covariates. Although the most efficient estimator in each class is often impractical, an estimator that estimates the sampling probabilities from the data achieves good efficiency in many realistic situations. This methodology allows for oversampling the most informative cohort members while exploiting information collected on all cohort members. Our R package EfficientCohort implements this methodology to produce survival curves and attributable risks adjusted for confounders. We demonstrate EfficientCohort on a cohort study of esophageal cancer and zinc, where measuring zinc requires precious esophageal biopsy tissue and thus is observed only on 25% of the cohort members.

email: katkih@mail.nih.gov

AN ADAPTIVE BAYESIAN DESIGN FOR A RANDOMIZED TRIAL OF HIGH- VERSUS LOW-LEVEL TACROLIMUS AS PROPHYLAXIS FOR GRAFT VERSUS HOST DISEASE IN ALLOGENEIC STEM CELL TRANSPLANTATION

Peter F. Thall, University of Texas M. D. Anderson Cancer Center
Leiko H. Wooten*, University of Texas M. D. Anderson Cancer Center
Daniel R. Couriel, University of Texas M. D. Anderson Cancer Center
Richard E. Champlin, University of Texas M. D. Anderson Cancer Center

We describe a Bayesian adaptive design for a randomized phase II/III clinical trial to compare two different tacrolimus blood levels for the prevention of GVHD after allogeneic hematopoietic stem cell transplant. The scientific objectives are to estimate differences in overall survival, regimen-related mortality, acute GVHD, and toxicity at high versus low tacrolimus blood levels. The trial currently is ongoing at M.D. Anderson Cancer Center. The clinical goal of the trial is to find the lowest blood level of tacrolimus that can give the highest amount of protection against GVHD with the fewest side effects possible. A Bayesian probability model accounting for patient prognostic covariates (age and cell source) and the times to acute GVHD, disease recurrence, and death from time of transplant was used as a basis for constructing multiple interim monitoring rules to terminate either treatment arm for safety or stop the trial due to either futility or superiority. The rules are applied continuously throughout the trial, with either arm stopped early if the probability of (1) acute GVHD within 100 days or (2) or disease recurrence or death within 180 days is unacceptably high. Additionally, the trial will be stopped early if the posterior probability that one arm is superior to the other in terms of their 180-day success rates exceeds 99.5%.

email: leiko@mdanderson.org

ABSOLUTE EVENT RATES AND COX REGRESSION

James B. Kampert*, Cooper Institute

In many epidemiological investigations relating risk of an event to exposure, results of Cox proportional hazards regression analysis are graphically presented as hazard ratios for various levels of exposure compared with a reference level or category. Although this reveals the dose response relation between risk and exposure, absolute event rates are not apparent, and confidence intervals are tied to an arbitrary reference exposure level, where the interval vanishes. In prospective follow-up studies, however, absolute event rates and their accuracy are both meaningful and interesting across all levels of exposure. We present a method to determine absolute event rates expressed as events per unit of observation time at each level of exposure after adjustment for covariates, and we obtain pointwise standard errors and confidence intervals. These estimates only require quantities routinely derived from a Cox regression analysis by exploiting an analogy with exponential survival models. We illustrate with prospective data relating risk of all-cause mortality to cardiovascular fitness as measured by a maximal treadmill test at the Cooper Clinic in Dallas, TX, during 1970 to 1998. We also present simulation studies to evaluate the performance of the method. Supported by NIH/NIA AG06945

email: jkampert@cooperinst.org

A GENERAL FAMILY OF CURE MODELS

Ning Liu*, University of Michigan
Jeremy Taylor, University of Michigan

Cure models in survival analysis are useful in situations where it is reasonable to assume that a fraction of the observations will not experience the event of interest even if followed for a long time. This occurs when considering recurrence of cancer following treatment and in many other applications. There are two formulations of a cure model, one as a mixture model where the population is a mixture of cured and non cured subjects, while in the other formulation the cumulative hazard is assumed to be bounded. In both models covariates can affect both the cured fraction and the distribution of event times. In this paper we propose a family of cure models, indexed by a Box-Cox type transformation parameter, such that each of the formulations of the cure model are special cases. Numerical algorithms are developed to maximize the likelihood. Simulation studies are presented and a head and neck cancer dataset is discussed.

email: liuning@umich.edu

USING LOCALLY WEIGHTED REGRESSION MODELS TO ESTIMATE FAMILIAL CORRELATIONS

Zhiying Xu*, Case Western Reserve University
Qing Lu, Case Western Reserve University
Robert C. Elston, Case Western Reserve University
Sudha Iyengar, Case Western Reserve University

Family data are often collected to study the genetic mechanism underlying a disease trait. Familial correlations are estimated to quantify the heritability as a measure of the potential genetic contribution available in the ascertained families. In many cases the distributions of the quantitative traits for parents and offspring are measured on different scales due to biological causes. In this paper, we propose a two-step strategy to estimate the heritability of a disease trait for data where parents and offspring are at different developmental stages. In the first step, a robust locally weighted regression model (e.g. a LOWESS spline) is chosen to adjust for covariates that confound the trait values. We then use residuals from this nonlinear spline fit to estimate, in the second step, familial correlations. During the first step, most applications assume a linear or simple polynomial relationship between the trait and the covariates, with underlying bivariate normality of the residuals between relationship types. These assumptions may reduce familial correlations, if the relatives' residuals are not linearly correlated, and reduce estimates of heritability. We applied our strategy to estimate parent-offspring correlations, and hence heritabilities, for sound-speech disorder data.

email: zxu@darwin.cwru.edu

GLUMIP 2.0: FREE SAS/IML® SOFTWARE FOR PLANNING INTERNAL PILOTS

John A. Kairalla*, University of North Carolina at Chapel Hill
Christopher S. Coffey, University of Alabama at Birmingham
Keith E. Muller, University of North Carolina at Chapel Hill

We present the latest version of our free SAS/IML® software for planning internal pilot studies. Internal pilot designs involve conducting an interim power analysis (without interim data analysis) to modify the final sample size. For internal pilot studies in the General Linear Univariate Model (GLUM) framework, an unadjusted hypothesis test may lead to test size inflation. A 'bounding' test achieves control of test size while still providing most advantages of the unadjusted test. However, computational tools utilized in the previous version of the GLUMIP software for this method were very slow and unstable. Our new exact theory, based on simple forms for the test statistic density and derivative needed for the 'bounding' method is much more fast and stable and still works for the broad classes of ANOVA and regression problems handled by the old version. The new analytic forms incorporated into the software solve many problems inherent to current internal pilot techniques for linear models with Gaussian errors. Hence, the GLUMIP 2.0 software makes it easy to perform exact power analysis for internal pilots under the GLUM framework with Gaussian errors. Restrictions include: assumption of Gaussian errors, fixed predictor values, common design for all replications, and no missing data.

email: jkairall@bios.unc.edu

ON SIMULATIONS OF META-ANALYSIS OF LINKAGE RESULTS

Weihua Guan*, University of Michigan
Michael Boehnke, University of Michigan

We are comparing statistical methods of meta-analysis for human genetic linkage studies. This work is inspired by a meta-analysis study we are conducting for the International Type 2 Diabetes Linkage Analysis Consortium using the genome scan meta-analysis (GSMA) method of Wise and colleagues (1999). GSMA divides the genome into bins of ~30 cM, ranks the best linkage results in the bins for each sample, and then sums the ranks across studies; bins with high ranks are suggested as likely locations of disease-predisposing variants. We have carried out analyses of the diabetes data and of simulated data in which we used ranks, or instead used maximum LOD scores, minimum p-values, or truncated minimum p-values (Zaykin et al. 2002), and used either 30cM or 2cM bins. Our results suggest that: (1) under the ideal condition where markers are evenly spaced, all methods perform well; (2) using a single genetic map across all samples, the type-I error rate is inflated in 30cM-bin methods when markers are densely spaced; (3) using sample-specific maps, all methods have an inflated type-I error rate. Among these methods, the LOD-score and truncated p-value methods perform somewhat better than the others.

email: wguan@umich.edu

MISSING DATA IN MULTIVARIABLE RISK ADJUSTMENT MODELS

Paul Kolm*, Emory University School of Medicine
Emir Veledar, Emory University School of Medicine
Jovonne K. Foster, Emory University School of Medicine
Kathleen Hewitt, American College of Cardiology
Kristi R. Mitchell, American College of Cardiology
Richard E. Shaw, Sutter Pacific Heart Centers
Ralph E. Shaw, Sutter Pacific Heart Centers
Ralph G. Brindis, Kaiser Permanente, CA
Viola G. Brindis, Kaiser Permanente, CA
Viola Vaccarino, Emory University School of Medicine
William S. Weintraub, Emory University School of Medicine

Development of risk adjustment models of adverse outcomes in cardiovascular patients is often plagued by missing data for one or more of the risk factors considered for inclusion in the model. Multiple imputation methods have been developed to allow all patients to be included in the analysis. These methods assume that missing data are missing at random (MAR). Acute MI patients with elevated ST segment undergoing PCI were selected from the American College of Cardiology-National Cardiovascular Data Registry (ACC-NCDR) for analysis of risk factors of in-hospital mortality. Two of the risk factors considered, ejection fraction (EF) and body mass index (BMI), were missing for some of the patients. Acute MI patients with multiple complications, and more likely to experience the adverse outcome, may not be subjected to EF procedures by physician choice; thus missing EF values are unlikely to be MAR. Classification trees of in-hospital mortality were developed that included missing as another factor. The classification tree grouped missing EF with EF < 40% at all levels of the tree in which EF appeared. Missing BMI was not associated with in-hospital mortality. Tree-based models that include missing as another factor may provide quantitative evidence of MAR.

email: paul.kolm@emory.edu

IMPROVING ESTIMATION VIA DEPENDENT CENSORING

Chris A. Andrews*, Oberlin College

It is common for sacrifice times to be specified prior to the start of an experiment. This prevents a dependence between the response variable and monitoring time, which would bias most estimators. However, more precise information can be gathered if sacrifice times are allowed to depend on time dependent covariate processes that are monitored during the experiment. The locally efficient estimation procedure of Andrews, van der Laan, and Robins remain consistent in the presence of dependent censoring. The context of this talk is the estimation of the dependence of the time of tumor onset in mice on drug dosage and other covariates. Current status data (interval censored data, type I) on the time of tumor onset is observed. Weight data is collected and used to determine sacrifice time.

email: chris.andrews@oberlin.edu

SURVEY RESPONDERS, NONRESPONDERS, AND ACTIVE REFUSERS IN A STUDY OF CHRONIC PAIN

Peter C. Wollan*,Olmsted Medical Center
Emmeline R. Watkins, AstraZeneca
Barbara P. Yawn, Olmsted Medical Center

In the Spring of 2004, a questionnaire about chronic pain was mailed to a random sample of residents of Olmsted County in Southeast Minnesota. Subjects were also asked to sign HIPAA permission forms to allow access to their medical records. 3531 subjects completed all forms (responders); 681 explicitly declined to complete the questionnaire (questionnaire refusers); 45 completed the questionnaire but refused access to their medical records (HIPAA refusers); and 1642 did not respond (non- responders). In accordance with Minnesota law, for those subjects who did not respond, and who had not refused general research access at the institution where they received medical care, access to medical records was permitted. Visit and diagnosis data were retrieved from billing records for 5805 subjects. We compare responders, questionnaire refusers, and non-responders, using information from medical records, and we compare the HIPAA refusers to the responders using information within the questionnaire. Preliminary results indicate substantial differences among all four groups.

email: pwollan@olmmed.org

A MODEL FOR HIGHLY SKEWED AND ROUNDED REPEATED MEASURES DATA

Huichao Chen*, Emory University
Amita K. Manatunga, Emory University
Robert H. Lyles, Emory University

A random effects Weibull model is proposed for highly skewed and rounded exposure data. Rounding frequently occurs when exposures are determined via laboratory methods, and can lead to a combination of 'coarse' and left- censored data. Reported values fail to accurately reflect the actual knowledge about each measurement, and often appear to reflect a complex mixture of ordinal and continuous data. The proposed modeling approach is based on the assumption that the observed rounded exposure is determined by the value of an underlying unobservable continuous response that follows a Weibull distribution with random effects. For the Hougaard family frailty distributions, including the gamma distribution often employed in the literature, the resulting marginal likelihood obtains a closed form without resorting to numerical or approximation methods. The performance of the proposed model is supported by simulation studies and its application is illustrated using repeated polybrominated biphenyl (PBB) exposure data on participants of the Michigan Female Health Study.

email: hchen4@sph.emory.edu

A BAYESIAN PARAMETRIC TOOL TO ASSESS NORMALITY

Ilsung Chang*, Texas Transportation Institute

Box and Tiao (1973) presented a methodology how to confront the abnormality in the data. As a tool of Bayesian assessment of normality assumptions they suggested several parametric families that extend the normal family, one of which is the family including power transformation parameter. The power transformation has been widely used, especially, to incorporate the skewness. The posterior distribution of the additional parameter, the parameter of power in the transformation, could be important for the decision making. An alternative to the Box-Cox transformation will be presented as assessment of normality in the case where skewness exists in the data. The Azzalini model (1985) will be considered to cope with the skewness by allowing an additional parameter to the model. The posterior distribution of the skewness parameter is the key to the inference, which will be obtained without any MCMC technique since a normal prior for the skewness parameter makes possible the analytic integration for the posterior distribution. We will compare its performance to several classical normality test including Wilk-Shapiro test statistic.

email: ischang@stat.tamu.edu

VARIANCE COMPONENT ESTIMATION USING THE ADAA MODEL WHEN GENETIC DESIGNS ARE PARTIAL AND COMPLETE

Jixiang Wu*, Mississippi State University
Johnie N. Jenkins, Mississippi State University
Jack C. McCarty, Mississippi State University
Dongfeng Wu, Mississippi State University

In addition to additive and dominant genetic effects, the additive by additive interaction (AA epistatic) effects that control many quantitative traits are important for genetic and breeding studies. Estimation of these genetic variance components including G¡ÁE interaction requires data containing at least parents and other two generations (i.e. F1 and F2) in a complete experimental design. Practical difficulties may arise in running a complete design. We compared the estimated variance components among six partial and complete designs through simulation. These designs were as follows, Design 1: parents, F1s and F2s in two environments; Design 2: parents, F2s and F3s in two environments; Design 3: parents and F1s in environment 1 and parents and F2s in environment 2; Design 4: parents and F2s in environment 1 and parents and F3s in environment 2; Design 5: parents and F1s in environment 1 and parents, and F1s and F2s in environment 2; and Design 6: parents and F2s in environment 1 and parents, F2s and F3s in environment 2. Designs 5 and 6 gave similar mean estimated variance components with a little larger mean square error. Designs 3 and 4 would have the risk to obtain biased estimation for dominant variance component.

email: jw7@ra.msstate.edu

ORDER RESTRICTED INFERENCE APPROACH TO GENE EXPRESSION (ORIGEN)

Eric Harvey*, Constella Health Sciences
John Zajd, Constella Health Sciences
Shawn Harris, Constella Health Sciences
Joel Parker, Constella Health Sciences
Shyamal Peddada, National Institute of Environmental Health Sciences

ORIGEN is a user friendly Java-based software package for selecting and clustering genes according to their time-course or dose-response profiles. It is based on the methodology developed in Peddada et al. (Bioinformatics, 2003). This method takes into account the inherent structure present in time course and dose response microarray experiments. The user pre-specifies the profiles of interest. Using a bootstrap procedure and user specified level of significance, the ORIGEN program selects significant genes and clusters them into the best fitting profiles. The results for significant genes are saved to a text file and displayed graphically. Gene ontology information for significant genes is provided where available. ORIGEN is freely available through an NIEHS website.

email: rclee@constellagroup.com

STOCHASTIC SEARCH VARIABLE SELECTION FOR THE DETECTION OF DIFFERENTIAL GENE EXPRESSION

David A. Henderson*, University of Arizona

Statistical investigations of treatment contrasts within gene are often employed following adjustment for known sources of extraneous variation in microarray experiments regardless of array platform. Typically, these investigations inspect all possible within gene contrasts. Thus significance thresholds must be adjusted to account for the multiple hypotheses tested, resulting in decreased power for the detection of differential expression. Additionally, prior beliefs are that only a small subset, say less than 10%, of the genes are truly differentially expressed within a given experiment. The implementation of a stochastic search variable selection (SSVS) on a set of transformed variables, the linear treatment contrasts within gene, following adjustment for nuisance factors in gene expression experiments may result in increased power to detect differential gene expression and clarify interpretation. SSVS also generates posterior probabilities of models, sets of contrasts that collectively explain a significant proportion of the observed variance in gene expression, which may elaborate upon novel genetic mechanisms relevant to the experiment at hand. The increased power and efficiency of SSVS over more traditional methods for detecting differential gene expression will be demonstrated using a subsets of genes from previously analyzed array experiments.

email: dnadave@u.arizona.edu

MIXTURE MODEL FOR CLUSTER VALIDATION IN A GENE EXPRESSION MICROARRAY STUDY

Carrie M. Nielson*, University of Arizona
David A. Henderson, University of Arizona
George Watts, University of Arizona

A mixture model is applied to microarray data for 11 normal and 18 Barrett's esophagus samples to determine if a portion of the Barrett's esophageal samples are contaminated by normal cells. Previous hierarchical clustering results have indicated three distributions, two distinct Barrett's subgroups and a cluster of normal samples. A result of three distributions would validate hierarchical clustering results and may explain differences in cancer prognoses for Barrett's esophagus. Methods: 444 genes' expression values were previously determined to be statistically significantly differentially expressed among normal, Barrett's, and esophageal cancer biopsy samples using linear modeling in R. Only those values for normal and Barrett's esophagus samples were used in this analysis. Likelihoods were estimated for mixtures of two and three normal distributions as well as for a single normal distribution. A likelihood ratio test was performed between all three comparisons. Results: A likelihood ratio test of three vs two, three vs one, and two vs one distribution(s) yielded a p value for a chi-square of $p>0.05$. Conclusions: Results of a mixture model analysis on the gene expression values from normal and Barrett's esophagus biopsy samples do not support the division of Barrett's esophagus samples into either three or two separate clusters as was found by hierarchical clustering.

email: cnielson@email.arizona.edu

GENERALIZED SPATIAL STRUCTURAL EQUATION MODELS

Xuan Liu *, University of Minnesota
Melanie M. Wall, University of Minnesota
James S. Hodges, University of Minnesota

It is common in public health research to have high dimensional, multivariate, spatially-referenced data representing summaries of geographic regions. Often it is desirable to examine relationships among these variables both within and across regions. An existing modeling technique called spatial factor analysis has been used and assumes that a common spatial factor underlies all the variables and causes them to be related to one another. An extension of this technique considers that there may be more than one underlying factor, and that relationships among the underlying latent variables are of primary interest. However, due to the complicated nature of the covariance structure of this type of data, existing methods are not satisfactory. We thus propose a generalized spatial structural equation model (GSSEM). In the first level of the model, we assume the observed variables are related to particular underlying factor. In the second level of the model, we use the structural equation method to model the relationship among the underlying factors and use parametric spatial distributions on the covariance structure of the underlying factors. We apply the model to county-level cancer mortality and census summary data for Minnesota, including socioeconomic status and access to public utilities.

email: xuanliu@biostat.umn.edu

NOTES

ENAR WEBSITE
WWW.ENAR.ORG

ENAR

Index of Participants

ENAR

INDEX OF PARTICIPANTS

ENAR

Index of Participants

Index of Participants

Index of Participants

Index of Participants

Index of Participants

ENAR

INDEX OF PARTICIPANTS

Index of Participants

ENAR

Index of Participants

Index of Participants

Index of Participants

NOTES

ENAR WEBSITE
WWW.ENAR.ORG

NOTES

ENAR WEBSITE
WWW.ENAR.ORG

Congratulations to ENAR on the 2005 Spring Meeting!

The International Biometric Society (IBS) welcomes you to Austin, Texas for the 2005 ENAR Spring Meeting.

The IBS is an international, professional Society devoted to the development and application of statistical and mathematical theory and methods in the biosciences. The strength of the IBS is its member Regions and National Groups. There are over 25 IBS Regions and National Groups with ENAR being the largest Region. As a member of ENAR, you are also a member of the IBS and entitled to the following IBS benefits:

- Free subscription to Biometrics, our flagship journal. This includes a printed copy AND online access with searchable content (current volume and previous five volumes). There are four issues each year.

- Free access to Biometric Bulletin online and one copy of the annual print edition of the newsletter. There are three online issues and one special printed issue each year.

- Access to Biometrics Volume 1/1945 through 54/1998 for a nominal fee

- Reduced fees for attending the International Biometric Conferences (upcoming IBC2006 will be in Montréal, Canada, 16-21 July 2006. IBC2008 is in Dublin, Ireland!)

- Access to all active members contact information through the online directory

Discounts to the following journals:

- Biometrical Journal www.biometrical-journal.de. As an IBS member, you can receive a special discount of 12 % off the regular personal subscription price. The regular price is 228 USD (plus VAT and shipping costs). The special offer is for 6 issues at 184 USD or 136 Euro per subscription. Please contact Marketing Manager Veronique Bluteau (vbluteau@wiley-vch.de) for further details. Be sure to mention that you are an IBS member.

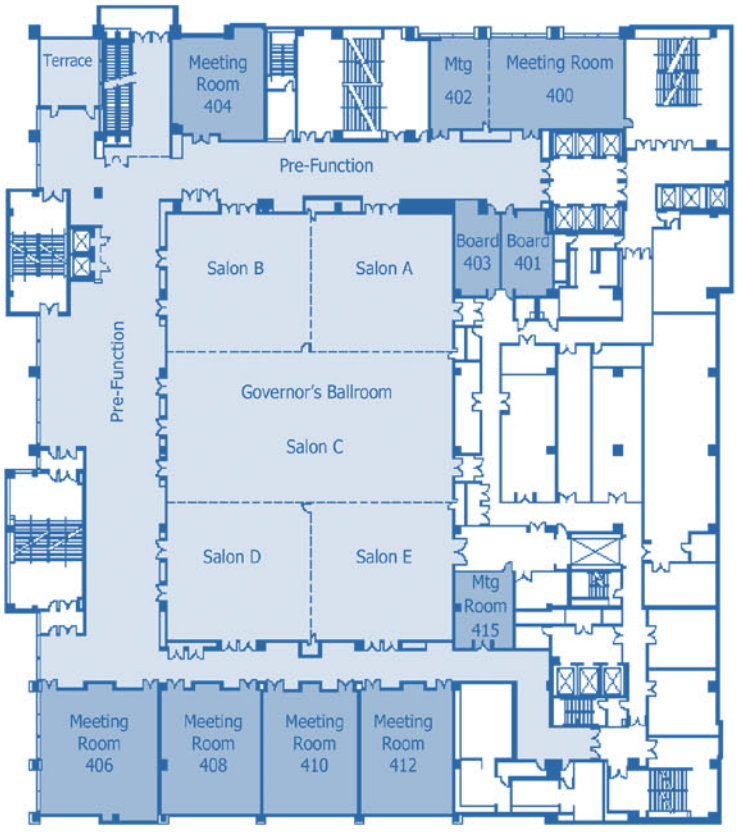
- Statistical Modelling *www.statmod.com*
  (US) $126.00 (normally $140.00)

- Statistical Methods in Medical Research *www.smmrjournal.com*
  (US) $130.00 (normally $172.00)

To access all of the benefits, please go to *www.tibs.org* and click on the‘members only’ section.

ENAR

HILTON AUSTIN CONVENTION CENTER HOTEL
FOURTH FLOOR



HILTON AUSTIN CONVENTION CENTER HOTEL
SIXTH FLOOR